UQAC
Université du Québec
à Chicoutimi

Exploration d'articles scientifiques sur les maladies rares pour l'extraction d'informations

Par

Charles Cousyn

Mémoire présenté à l'Université du Québec à Chicoutimi en vue de l'obtention du grade de

Maître ès sciences en Maîtrise en Informatique profil recherche

Québec, Canada

RÉSUMÉ

Les maladies rares constituent un sujet peu connu du grand public. Néanmoins, malgré leur nom, un grand nombre de personnes sont affligées par une ou plusieurs d'entre elles. La recherche sur près de sept mille maladies rares est insuffisante, et même si certains travaux ont été réalisés pour exploiter les publications scientifiques et extraire des informations pertinentes, les connaissances sont très difficiles à obtenir pour la population en général.

Ce document présente un nouveau système qui tente d'aborder l'extraction des connaissances sur les maladies rares dans les publications scientifiques. En particulier, nous nous concentrons sur la tâche d'extraire automatiquement les symptômes de maladies rares à partir de publications avec une nouvelle approche utilisant un algorithme de reconnaissance d'entité nommée (NER) basé sur la statistique numérique Term Frequency - Inverse Document Frequency (TF-IDF). L'approche envisagée permet d'atteindre un F-score de 17.17% avec une évaluation sur près de 3000 maladies rares, ce qui représente un progrès dans le cadre de l'extraction de symptômes de maladies rares à partir de publications scientifiques.

Ce mémoire est séparé comme suit. Le premier chapitre introduira le contexte, les motivations, le problème de recherche, les contributions et la méthodologie. Le second chapitre est une revue de littérature qui présentera les travaux pertinents de ce travail de recherche et permettra de définir la direction prise par ce projet par rapport au sujet des maladies rares. Le troisième chapitre introduira les étapes, les concepts ainsi que les termes importants à définir dans le cadre d'un projet de « text mining ». Le quatrième chapitre décrira les différentes contributions du travail de recherche en précisant les éléments extraits, les sources de données et les algorithmes utilisés (TF-IDF modifié), sans oublier la description de l'outil développé et la phase d'expérimentation. Enfin, le dernier chapitre conclura ce travail de recherche par une revue des contributions, les limites du travail de recherche, les travaux futurs envisageables et une conclusion personnelle sur le projet de recherche.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

NER:	NAMED ENTITY RECOGNITION
PMC:	PUBMED CENTRAL
SVM:	SUPPORT VECTOR MACHINE
NLP:	NATURAL LANGUAGE PROCESSING
XML:	EXTENSIBLE MARKUP LANGUAGE
JSON:	JAVASCRIPT OBJECT NOTATION
NLM:	NATIONAL LIBRARY OF MEDICINE
HPO:	HUMAN PHENOTYPE ONTOLOGY
CNN:	CONVOLUTIONAL NEURAL NETWORK

REMERCIEMENTS

Ce projet a été effectué au sein du laboratoire LIARA (Laboratoire d'Intelligence ambiante pour la Reconnaissance d'Activités) à l'UQAC (Université du Québec à Chicoutimi). Je tiens à remercier le laboratoire pour l'opportunité qui s'est présentée à moi.

Désireux de découvrir le domaine de la recherche scientifique, mes différents échanges avec Kevin Bouchard en tant que professeur pour le cours d'Intelligence artificielle m'ont fait comprendre que les domaines de l'intelligence artificielle et le « data mining » étaient des domaines pour lesquelles j'ai un intérêt prononcé.

Je tiens à remercier particulièrement Kevin Bouchard pour sa supervision, l'écoute ainsi que la confiance qu'il a témoignée tout le long de ce travail de recherche.

Finalement, je souhaite remercier mes collègues de laboratoire pour les échanges que j'ai pu avoir avec eux, leurs conseils, quand les connaissances manquaient. J'ai découvert des gens passionnés avec qui je partage aujourd'hui une amitié et je leur souhaite la réussite dans tous les projets qu'ils pourront entreprendre.

CHAPITRE 1

INTRODUCTION

1.1 CONTEXTE ET MOTIVATIONS

1.1.1 CONTEXTE THÉORIQUE

La quantité de données qui a été produite par l'humanité est tout simplement gigantesque. Avec l'avancée des technologies de stockage, et plus généralement des technologies de stockage en ligne, que cela soit pour stocker des informations personnelles, pour avoir à portée de main du divertissement (musiques, films, jeux, etc.), pour partager du contenu avec d'autres personnes, les humains ne cessent de trouver de nouvelles façons de faire disparaître les supports « physiques » de stockage d'informations au profit de supports dématérialisés moins spacieux, plus économiques et écologiques. Avec l'arrivée d'internet, le nombre de données existantes a été décuplé, et ce dans tous les domaines : biomédical, chimie, physique, économie, social, etc.

Nous avons donc, à notre disposition, une énorme masse de données. L'analyse de celles-ci peut rendre de nombreux services et permettre de grandes avancées scientifiques et technologiques. Un des exemples les plus récents de cette utilisation de données est le projet ImageNet par Deng et al. (2009). Ce projet consiste en la création d'une base de données de 15 millions de photos, chacune d'entre elles étant étiquetée avec les différents objets qui se trouvent sur celle-ci dans le but de faire de la reconnaissance d'image par ordinateur. En 2016, un « Convolutional Neural Network » (CNN) a ensuite été appliqué à toutes ces photos pour apprendre l'ordinateur à reconnaître les objets. Bien qu'on soit arrivé à des prouesses comme de la reconnaissance d'image fiable, il se trouve qu'encore aujourd'hui, les données disponibles restent trop peu utilisées.

Cette sous-exploitation n'est pas sans raison, une des premières raisons est qu'il n'est pas toujours facile d'exploiter les données. Premièrement, l'exploitation de données personnelles peut menacer la vie privée des utilisateurs. Le risque est qu'au lieu de simplement répondre à une question métier (c'est-à-dire, qu'elles servent à résoudre un problème), les données soient utilisées à une

autre fin. Ce détournement des données engendre souvent des problèmes tels qu'une atteinte à la vie privée, vol d'informations bancaires, usurpation d'identité, etc. Ensuite, l'interopérabilité des systèmes est un point capital pour permettre d'exploiter des données provenant de sources hétérogènes (base de données, fichiers, capteurs, etc.). Enfin, il est souvent nécessaire de disposer d'une certaine qualité de données, c'est-à-dire qu'elles doivent être cohérentes, pertinentes, et les plus complètes possible pour une bonne exploitation. C'est une des raisons principales pour lesquelles l'extraction de données peut être une tâche difficile.

Malgré ces points de blocages, de nombreuses applications utilisant l'exploitation de données ont vu le jour grâce aux avancées scientifiques qui ont été effectuées. Parmi ces avancées, on en retrouve dans le domaine des statistiques. Dans son livre, « Exploratory Data Analysis », Tukey (1977) énonce que les statistiques exploratoires permettent, en explorant les données, d'avoir une idée experte du fonctionnement du système qu'elles représentent. Cela permet la formulation d'hypothèses cognitives sur les phénomènes mis en jeu de leurs propriétés. En fin de compte, elles permettent de décrire le comportement des données sous un certain angle. Par exemple, en statistiques, la régression linéaire est un modèle de régression qui cherche à établir une relation linéaire entre une variable, dite expliquée, et une ou plusieurs variables, dites explicatives. Il existe aussi d'autres approches comme le « data mining ». Cette approche consiste en l'extraction d'un savoir ou d'une connaissance à partir d'une grande quantité de données par des méthodes automatiques. Elle se repose sur des méthodes statistiques et l'intelligence artificielle pour construire des modèles à partir des données.

Un cas particulier du « data mining » est le « text mining ». Le « text mining » est en fait l'utilisation de méthodes de « data mining » pour découvrir des motifs intéressants à partir de texte. L'utilisation de texte fait que le « text mining » repose sur l'exploitation de données non structurée, ce qui représente un challenge supplémentaire. Le « data mining » et le « text mining » peuvent souvent se décrire comme une suite d'étapes sur les données : nettoyage, intégration, sélection, transformation et exploitation de données, évaluation des motifs extraits et leur représentation. Toutes ces étapes sont importantes pour extraire des connaissances cohérentes et utilisables.

1.1.2 CONTEXTE APPLICATIF, DE LA SANTÉ, DU BIOMÉDICAL

Ce travail de recherche s'inscrit dans le contexte applicatif de la santé. En effet, la santé est une grande préoccupation de la communauté des chercheurs. Beaucoup de progrès ont vu le jour, de par l'amélioration des conditions, du confort de vie, des traitements des maladies, de la qualité des diagnostics dans les pays industrialisés. Malgré tout, dans ce contexte, de nombreux aspects de la santé restent peu étudiés, c'est le cas notamment des maladies rares connues. On dénombre entre 6000 et 7000 maladies rares, leur nombre dépendant de la source et des critères d'inclusion utilisés. Dans le cas de ces maladies, il est difficile d'établir un diagnostic fiable pour plusieurs raisons : le nombre de spécialistes d'une maladie rare est restreint, l'accès à ces spécialistes est limité et il existe un manque de connaissances et d'informations scientifiques à propos de la plupart des maladies rares. Pourtant, on estime qu'une personne sur 2000 serait affectée par une maladie rare dans la population de l'Union européenne, soit au moins 255 900 personnes (selon le rapport d'Orphanet (2016)).

Grâce à la disponibilité de données de littérature scientifique sur les maladies rares, il est possible d'envisager l'exploitation de techniques d'apprentissage machine dans le but d'automatiquement extraire des informations de qualité sur ces maladies afin d'aider les experts médicaux, ou les personnes atteintes, à trouver le bon diagnostic. Il serait aussi possible d'extraire de nouvelles hypothèses de recherche en mélangeant les écrits sur une maladie rare. Si deux articles, abordant la même maladie rare, font le lien entre la maladie et un certain gène (différent dans chaque article), il serait intéressant de signaler, à travers du « text mining » par exemple, qu'il existe potentiellement une relation entre ces deux gènes, et ainsi, ouvrir la porte à une nouvelle recherche cherchant à démontrer ou non l'existence de relation entre ces deux gènes.

1.1.3 DÉFIS EN RAPPORT AVEC LES MALADIES RARES

L'exploitation des publications scientifiques sur les maladies rares présente différents challenges à relever. Premièrement, les maladies rares, par leur nature, sont des maladies qui ont été moins sujettes à des recherches biomédicales et nécessairement, par rapport à des maladies plus courantes, il est possible que le nombre de publications ne soit pas assez grand. Il faut donc vérifier

qu'il existe suffisamment de publications scientifiques en ce qui concerne ces maladies. La vérification de cette information sera discutée dans la section « Hypothèses de recherche » (1.2.2).

Deuxièmement, même si le nombre de publications est suffisant, les publications scientifiques sont sous forme de texte et donc sous une forme non structurée que les méthodes statistiques, de « data mining » ne peuvent pas toujours exploiter au premier abord. Envisager d'apporter une structure cohérente au texte en fonction de ce qu'on veut extraire est important ; ce qui implique alors de réfléchir aux différents modèles de représentation que l'on pourrait utiliser pour structurer notre texte.

Enfin, il existe un dernier défi quant aux données disponibles actuellement sur les maladies rares. En effet, il existe plusieurs portails où l'on peut se renseigner sur les maladies rares. Ces portails sont construits par des personnes dévouées qui le font souvent gratuitement pendant leur temps libre ou par des organisations à but non lucratif travaillant à la sensibilisation à certaines de ces maladies (Orphanet, EpiRare, NORD, etc.). Cette stratégie d'information présente deux inconvénients majeurs. D'abord, le processus de transfert des connaissances débute par des spécialistes hautement qualifiés qui publient des papiers scientifiques, puis par les médecins qui utilisent ces publications pour agrandir leurs connaissances, puis par les personnes travaillant sur le portail. L'ensemble du processus peut alors prendre des années. Ensuite, au vu du nombre d'étapes du processus et au nombre d'humains pouvant être impliqués dans celui-ci, il y a des chances que l'information perde de son exactitude avec le temps.

1.2 PROBLÈME DE RECHERCHE

Au vu des défis que présentent les maladies rares, nous avons voulu orienter notre travail de recherche afin de les relever de la meilleure manière possible. Cette section s'occupera donc de présenter le problème de recherche. Nous commencerons par l'explication de la terminologie et des concepts du sujet de recherche, pour ensuite poser certaines hypothèses dont nous ferons la justification. Nous terminerons alors par l'énonciation de la question de recherche.

1.2.1 TERMINOLOGIES ET CONCEPTS

Ce travail de recherche concerne les maladies rares et les publications scientifiques, nous allons définir dans cette section ces deux notions dans cet ordre.

1.2.1.1 MALADIE RARE

Il existe différentes définitions de ce qu'est une maladie rare selon la localisation. Nous utiliserons la définition européenne, car elle est suffisamment vaste pour avoir assez de données à exploiter.

Selon le serveur d'informations sur les maladies rares, Orphanet (Inserm, 1999), les maladies dites « rares » sont celles qui touchent un nombre restreint de personnes et posent de ce fait des problèmes spécifiques liés à cette rareté. Le seuil admis en Europe est d'une personne atteinte sur 2 000. Une maladie peut être rare dans une région et fréquente dans une autre. Ainsi la thalassémie, une anémie d'origine génétique, est rare dans le nord de l'Europe, alors qu'elle est fréquente autour de la Méditerranée (Orphanet, 2015). Autre exemple, ce qu'on appelle une maladie périodique est considéré comme rare en France, mais se trouve être courant en Arménie. Il y a également des maladies fréquentes qui possèdent des variantes rares.

Les maladies rares se comptent certainement par milliers. À l'heure actuelle, on en a déjà dénombré entre 6000 et 7000 et de nouvelles maladies sont régulièrement décrites dans la littérature médicale. En ce qui concerne leurs origines, si presque toutes les maladies génétiques sont des maladies rares, toutes les maladies rares ne sont pas génétiques. Il y a des maladies infectieuses très rares par exemple, ainsi que des maladies auto-immunes et des cancers rares. Pour un grand nombre de ces maladies rares, la cause demeure inconnue à ce jour.

Les maladies rares souffrent d'un déficit de connaissances médicales et scientifiques. Elles ne sont apparues que récemment dans les politiques de recherche et de santé publique. Les personnes atteintes par ces maladies rencontrent toutes des difficultés similaires dans leur parcours vers un diagnostic, pour obtenir de l'information et pour être orientées vers les professionnels compétents. L'accès à des soins de qualité, la prise en charge globale sociale et médicale de la maladie, la coordination des soins hospitaliers et de ville, l'autonomie et l'insertion sociale, professionnelle et

citoyenne, posent également problème. Faute de connaissances scientifiques et médicales suffisantes, un grand nombre de malades ne sont pas diagnostiqués ; leurs maladies demeurent inconnues. Ces personnes sont alors prises en charge sur la base de l'expression de leurs symptômes.

1.2.1.2 PUBLICATIONS SCIENTIFIQUES

Afin de définir le concept de publication scientifique, on peut s'appuyer sur la définition donnée sur Wikipedia (2017) :

« L'expression publication scientifique regroupe plusieurs types de communications scientifiques et/ou techniques avancées que les chercheurs scientifiques font de leurs travaux en direction de leurs pairs et d'un public de spécialistes. Ces publications ayant subi une forme d'examen de la rigueur de la méthode scientifique employée pour ces travaux, comme l'examen par un comité de lecture indépendant constitué de pairs. »

Ce qu'il est important de retenir est que, dans ce type d'écrit, une vérification de la qualité de la méthode scientifique est effectuée. Cela permet notamment de s'assurer de la qualité des conclusions de recherche données, de leur véracité et ainsi permettre aux prochains chercheurs qui voudraient utiliser le contenu de l'article de s'appuyer sur des faits dont la véracité n'est pour l'instant pas remise en cause. Dans ce travail, le terme « publication » est important, car il existe différentes catégories de publications scientifiques :

- Les revues scientifiques à comité de lecture
- Les comptes-rendus de congrès scientifiques à comité de lecture
- Les ouvrages collectifs regroupant des articles de revue ou de recherche autour d'un thème donné
- Les monographies sur un thème de recherche

Ici, nous nous intéresserons bien aux publications scientifiques et non aux articles scientifiques seulement. Cela permettra d'étendre la quantité d'informations utilisables pour l'extraction de connaissances. Nous viserons à exploiter les publications focalisant sur une maladie particulière, mais due à la quantité de données, il ne sera pas possible de s'assurer qu'elles ne sont pas le sujet de multiples maladies et que certaines de ces maladies soient communes.

1.2.2 HYPOTHÈSES DE RECHERCHE

Le sujet de recherche étant vaste, nous avons défini et limité le champ de notre recherche en posant les hypothèses suivantes. La justification de ces hypothèses sera apportée dans cette section. Notons que certaines de ses hypothèses seront remises en question dans la section « Limites » (5.2).

1.2.2.1 HYPOTHÈSE H0

H0 : **Le moteur de recherche utilisé récupère des publications qui sont véritablement en lien avec les maladies rares lors d'une recherche.**

Pour obtenir des publications scientifiques, l'un des principaux moyens est d'utiliser un moteur de recherche en ligne. Par exemple, en tapant un terme dans la barre de recherche du site https://www.ncbi.nlm.nih.gov/pubmed/, on obtient une liste de résultats étant en lien avec notre terme de recherche. Nous posons l'hypothèse H0, car nous utiliserons un moteur de recherche utilisant les dernières avancées scientifiques pour donner des résultats dont nous détaillerons le fonctionnement dans la section 4.3.4. Il y a également le fait que si l'hypothèse H0 se trouve être non valide, il n'y a alors aucune raison justifiant que nous pourrions être capables de le faire nous-mêmes. Nous nous focaliserons donc principalement sur l'exploitation des publications scientifiques que le moteur de recherche nous fournira.

1.2.2.2 HYPOTHÈSE H1

H1 : **La grande majorité de l'information à propos d'une maladie rare se trouve dans les publications scientifiques en rapport avec cette maladie.**

La chaîne de la connaissance scientifique débute toujours par des travaux scientifiques qui seront retranscrits par écrit sous forme de publications, qui seront revues par les pairs, puis ces publications seront lues par d'autres scientifiques, les programmes d'enseignement ajouteront les connaissances provenant de ces publications et cet enseignement formera les nouveaux scientifiques. Ce projet souhaite aller chercher au plus tôt dans cette chaîne de la connaissance scientifique et l'un des maillons les plus facilement exploitables semble être les publications scientifiques.

Cette hypothèse H1 représente la base de ce mémoire, car le fait de la poser et de réaliser ce travail nous permettra de la tester et de savoir, si oui ou non, il y a de bonnes raisons de penser

qu'un programme informatique pourrait automatiquement extraire de l'information des écrits scientifiques. Plus spécifiquement, si elle s'avère infondée, nous ne pourrons simplement pas trouver l'information, puisqu'elle serait inexistante.

1.2.2.3 HYPOTHÈSE H2

H2 : **Les informations qui vont être utilisées pour évaluer notre extraction sont vraies.**

Évaluer l'outil implémenté est capital dans un projet de recherche. Dans notre cas, nous comparerons les données extraites automatiquement avec des données maintenues manuellement par une plateforme en ligne réputée sur le sujet des maladies rares (Inserm, 1999). Puisque cette plateforme représente l'information la plus fiable disponible au sujet des maladies rares, il n'y aurait pas de moyen de valider l'outil implémenté si H2 se révélait fausse.

1.2.2.4 HYPOTHÈSE H3

H3 : **Les éléments qui seront extraits par notre approche seront liés à une maladie précise**.

Par exemple, dans le cas de l'extraction d'anomalies phénotypiques (défini en 4.2.1, mais on peut ramener ce terme à la notion de « symptôme »), ces anomalies sont celles qu'une maladie précise provoque. Nous ne nous intéresserons pas à l'extraction d'anomalies phénotypiques de manière générale dans les publications scientifiques, mais plutôt à celles qui concernent chaque maladie. Nous voulons ainsi permettre, à partir du seul nom de la maladie, d'obtenir des éléments liés à celle-ci.

1.2.3 QUESTION DE RECHERCHE

Dans ce projet de recherche, l'objectif est d'aborder la question de l'extraction de connaissances sur les maladies rares en construisant un outil d'extraction exploitant les publications scientifiques par du « text mining ». La question de recherche qui sera traitée sera donc la suivante :

« Comment exploiter automatiquement les publications scientifiques sur les maladies rares pour extraire de la connaissance ? »

1.3 CONTRIBUTIONS

Pour répondre à cette question de recherche, ce mémoire de recherche a été réalisé en suivant une méthodologie scientifique classique. Il apporte une contribution à la discipline de la fouille automatique de textes à deux niveaux.

Au niveau théorique, ce mémoire tente d'introduire un nouvel algorithme de reconnaissance d'entités nommées pour les publications des maladies rares. Cet algorithme basé sur « TF-IDF »(Stephen, 2004) tente de montrer une nouvelle approche qui pourrait ouvrir la voie à un nouveau champ de recherche sur la reconnaissance d'entités nommées.

Au niveau pratique, la contribution se divise en trois parties. Premièrement, l'outil développé en C# sous Visual Studio propose un module agrégateur de publications scientifiques capable d'acheminer les articles scientifiques de la base de données PubMed Central (PMC) et de les stocker dans une base de données. La seconde contribution pratique est un module d'extraction d'entités des écrits scientifiques. L'outil utilise les algorithmes de reconnaissance d'entités nommées de la librairie LingPipe (Alias-i, 2008). Enfin, la dernière contribution est un module d'évaluation de la qualité de l'extraction.

1.4 MÉTHODOLOGIE

Tel que décrit précédemment, nous apportons notre contribution de recherche grâce à une méthodologie scientifique. Cette méthodologie s'est déroulée en trois étapes clés sur une période de 12 mois.

La première étape a tout d'abord été l'acquisition de connaissances dans le domaine en réalisant une revue de littérature sur l'extraction d'éléments dans les publications biomédicales avec des méthodes de « text mining ». Cette étape a permis de savoir quels sont les éléments qui peuvent être extraits des publications scientifiques par ces méthodes. Cela nous a aidés à connaître les problèmes bien connus de l'extraction ainsi qu'à faire notre choix dans les éléments à extraire.

La deuxième étape a été l'élaboration de l'outil d'extraction. Pour ce faire, l'outil a trois responsabilités distinctes. Dans un premier lieu, il doit s'occuper de la récupération et du stockage des publications scientifiques. La source des publications et de la méthode employée pour les récupérer et les stocker a beaucoup évolué depuis le début du projet et sera discuté dans la section « Sources

de données » (4.5). Deuxièmement, après avoir récupéré les publications, l'outil doit s'occuper de l'extraction des éléments du texte. Pour ce module, notre outil utilise une bibliothèque bien connue en « text mining » LingPipe (Alias-i, 2008) et plus particulièrement les méthodes de « named entity recognition » proposées. Enfin, une fois les éléments extraits, un module d'évaluation s'occupe de comparer les éléments extraits avec la vraie donnée en générant des fichiers avec de nombreuses mesures de performance.

La dernière étape est d'utiliser les fichiers générés afin de se faire une idée de la qualité de notre extraction et, ainsi de tenter d'améliorer au fur et à mesure cette dernière. Un outil de visualisation a été développé pour interpréter les fichiers d'évaluation ; il permet de faire des statistiques ainsi que des représentations graphiques des performances sous forme de graphes.

1.5 ORGANISATION DU DOCUMENT

Ce mémoire est organisé comme suit. Le présent chapitre introduit le problème de recherche de l'extraction d'informations sur les maladies rares à partir de publications scientifiques. Le deuxième chapitre est la revue de littérature des travaux scientifiques pertinents qui permettent de positionner ce travail de recherche au sein de la littérature scientifique. Plus spécifiquement, il permettra de bien justifier la valeur de la contribution scientifique. Le troisième chapitre introduira le sujet de l'analyse de texte et du « text mining », définira les concepts importants comme la phase de prétraitement, la représentation des données textuelles ainsi que les algorithmes utilisés couramment. Le quatrième chapitre présentera les contributions des travaux réalisés en posant la problématique, en expliquant les choix faits au niveau des sources de données, des algorithmes utilisés. Il contiendra également le bilan des expérimentations réalisées et l'interprétation qui en est faite. Finalement, le cinquième chapitre s'occupera de conclure ce travail de recherche en soulignant les contributions ainsi qu'en présentant les différentes limites et les travaux futurs de l'approche proposée. Ce chapitre se finira sur une conclusion personnelle de ce premier travail dans le monde de la recherche scientifique.

CHAPITRE 2

ÉTAT DE L'ART

L'extraction d'informations à partir de texte n'est pas quelque chose de nouveau (Schulman, Castellon, & Seligman, 1989), un grand nombre de méthodes existent et permettent de servir différents objectifs.

L'objectif de ce chapitre consiste à faire le tour des systèmes qui ont été développés et qui sont les plus pertinents par rapport à la problématique de ce travail de recherche. Nous commencerons par parler des critères d'inclusion, d'exclusion et d'analyse des travaux pour ensuite parler des travaux en eux-mêmes et nous finirons par une discussion et une conclusion de cet état de l'art.

2.1 CRITÈRES D'INCLUSION/EXCLUSION

Les travaux qui constituent l'état de l'art seront pris des bases de données scientifiques suivantes : Academic Search Complete, Computers & Applied Sciences Complete, ACM. Les publications sont obtenues, la plupart du temps en texte complet, grâce au service fourni par la bibliothèque de l'Université du Québec à Chicoutimi à l'adresse http ://libguides.uqac.ca/content.php ?pid=612816&sid=5064383. Les résultats des recherches seront limités aux travaux datant d'après 2012. Enfin, les mots-clés utilisés et plus particulièrement, la requête utilisée est :

"((statistic* AND (search* OR analysis)) OR text* mining OR machine learning OR extraction) AND (article* OR paper* OR publication* OR literature*)"

Quelques-uns des articles cités ont également été trouvés à l'aide du moteur de recherche Google Scholar (Google, 2004).

2.2 CRITÈRES D'ANALYSE

Les critères permettant d'analyser les travaux en fonction du besoin défini dans l'introduction sont :

- Modèle utilisé

- Connaissances extraites
- Performance
- Utilisation du texte des publications (complète, partielle)

La performance, elle, sera principalement basée sur le F-score, qui est une mesure permettant de représenter à la fois le taux de précision et le taux de rappel d'un modèle à évaluer. Plus concrètement, le F-score est un réel entre 0 et 1 qui sera considéré comme « bon » s'il est proche de 1 et « mauvais » s'il est proche de 0. Cette mesure ne dépend pas de la machine qui a fait l'évaluation, on pourra donc reprendre les F-score des travaux cités. Si le F-score n'est pas précisé, on prendra la valeur de l'amélioration donnée dans la publication. Le critère « connaissances extraites » définira le type d'informations extraites par les travaux effectués. Cela peut être des associations (relations), des faits, ou d'autres choses.

2.3 TRAVAUX

2.3.1 EXTENSION DE MOTEUR DE RECHERCHE

Dans leurs travaux, Cha, Kim, Yeu et Park (2016) ont tenté d'améliorer la quantité de résultats issus de recherches sur PubMed. PubMed est un moteur de recherche permettant de naviguer dans la base données MEDLINE. Ce genre d'outils se base souvent sur les titres de sujets médicaux et donc les résultats vont dépendre de l'indexation selon ces termes. Conséquemment, un certain nombre de données qui apporteraient de la valeur ajoutée sont ratées si l'indexation n'est pas effectuée adéquatement. Ainsi, par l'utilisation de « Support Vector Machines » (SVM) (Cortes & Vapnik, 1995), l'équipe est parvenue à augmenter de 40 % les résultats sur PubMed.

Ce gain d'information a son importance, car il permet de renforcer l'efficacité des chercheurs en biomédical, mais aussi des chercheurs en général. Dans ce travail, le modèle utilisé consiste en 5 étapes :

- Sélection des mots-clés et téléchargement
- Écrémage des informations (titre, PMID et résumé conservés)
- Prétraitement : conversion en données structurées

- Classification avec trois algorithmes : SVM, « Random Forest » et « Naives Bayes Classifier »
- Évaluation du modèle avec un ensemble de tests (créé manuellement au préalable)

2.3.2 EXTRACTION DES INTERACTIONS ENTRE PROTÉINES

Dans le travail de Li, Zang, Sun et Wang (2016), il a été possible d'extraire les interactions entre protéines des écrits scientifiques avec un F-score de 79,3 %. Ce travail est intéressant, car il met en évidence la possibilité d'extraire des relations entre des éléments, ici entre protéines, et cela reste proche de notre problématique. Les différentes étapes du processus sont les suivantes et sont représentées dans la Figure 1 :

- Collection de données : on extrait les résumés des publications scientifiques
- Prétraitement du texte : on supprime les duplicata, on identifie les protéines…
- Représentation structurée et extraction des caractéristiques
- Création du modèle et classification : ici, on construit le modèle avec différentes méthodes (SVM, Arbre de décision, « Random Forest », « Naives Bayes Classifier ») et on teste chacun des modèles

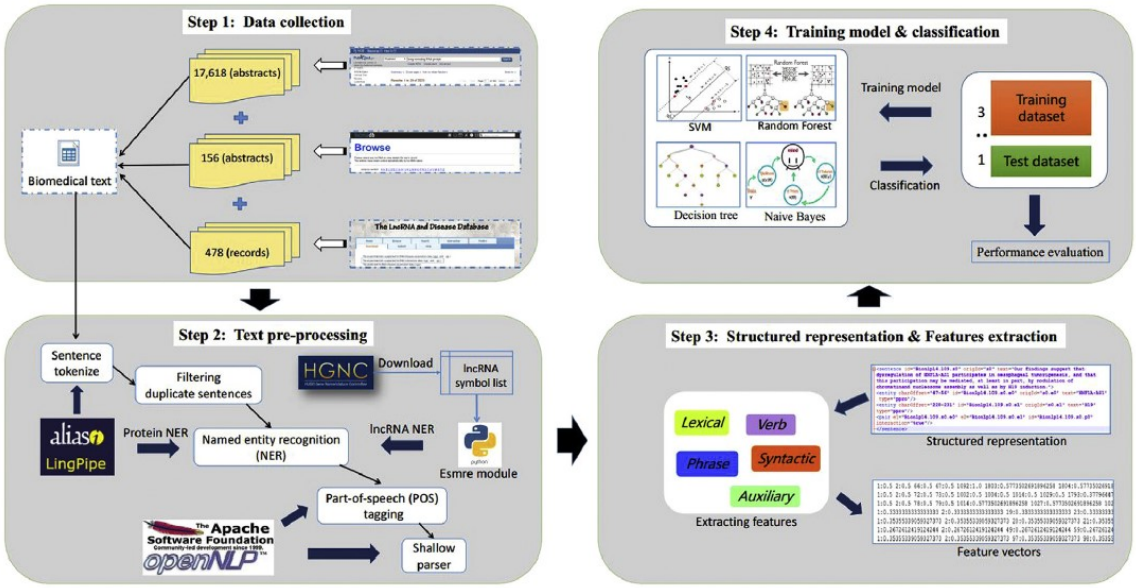

FIGURE 1: ÉTAPES DE LA MÉTHODOLOGIE PROPOSÉE

Finalement, l'algorithme ayant obtenu le meilleur score s'est avéré être « Random Forest », car il a dépassé les autres approches qui étaient « Support Vector Machines », « Naives Bayes Classifier » et « Decision tree ».

2.3.3 EXTRACTION D'ASSOCIATION MUTATION/MALADIE

Le travail de Mahmood, Wu, Mazumder et Vijay-Shanker (2016) se rapproche un peu plus de celui visé par ce mémoire de maîtrise. En effet, ce dernier, appelé DiMeX, permet l'extraction, à partir des résumés des publications scientifiques, d'associations entre mutation, maladies et gènes.

Pour ce faire, après avoir prélevé les informations sur les publications, différents éléments sont identifiés (maladies, gènes et contexte sur les patients). Une fois ces éléments identifiés, une analyse syntaxique est réalisée en identifiant les phrases nominales et les groupes verbaux. Ensuite, grâce à des expressions régulières, les mutations sont identifiées. Une fois tout cela réalisé, le système commence à faire des associations entre des éléments. D'abord, grâce à un principe de cooccurrence étendue, on fait des associations entre mutations et gènes, ensuite grâce à l'emploi de différents motifs de phrases d'association, on fait des associations entre mutations et maladies. Pour finir, un système de stockage est mis en place pour stocker toutes ces associations et des informations complémentaires avec un système de recherche pour effectuer des requêtes sur la base de données. L'architecture utilisée, représentée en Figure 2, est composée de 3 modules :

- Module d'identification (module A)
- Module de mise en relation (module B)
- Module d'informations complémentaires (module C)

Toutes les associations extraites stockées dans une base de données sont mises à disposition publiquement. Cette fois-ci, les F-score obtenus sur trois jeux de données différents sont de 88%, 91% et 89%.

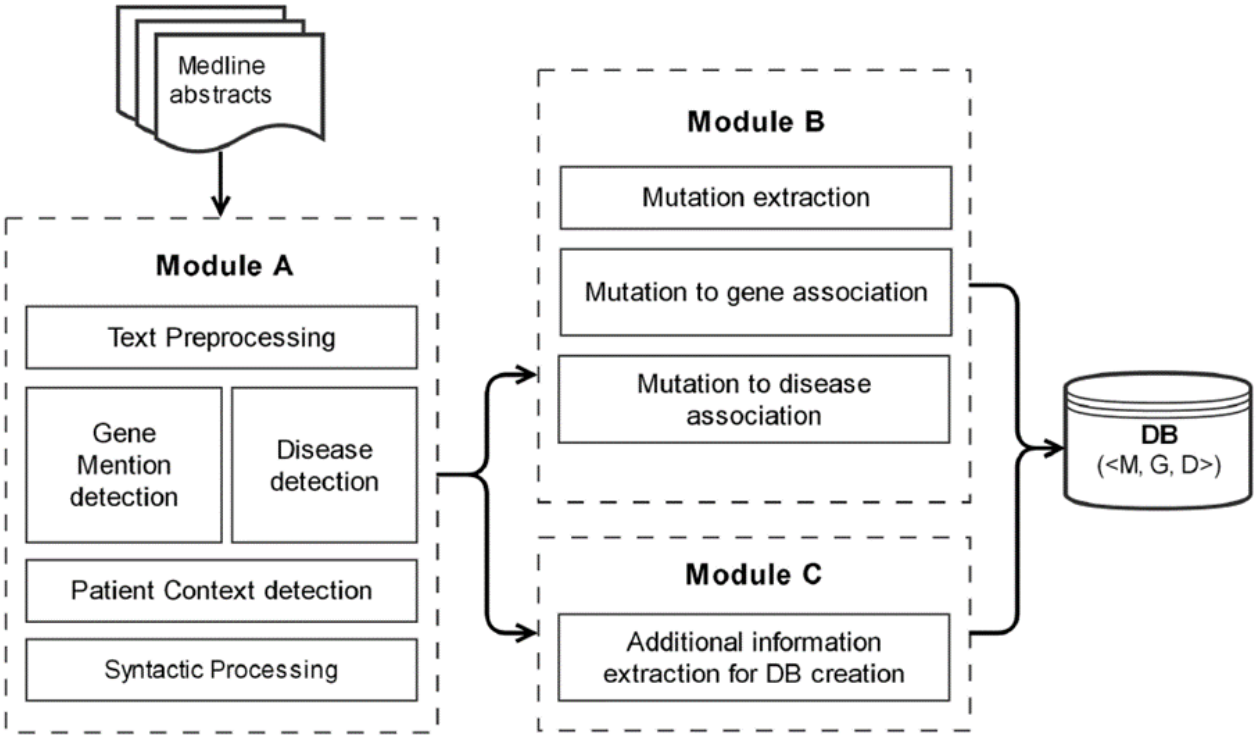

FIGURE 2: ARCHITECTURE DIMEX (MAHMOOD ET AL., 2016)

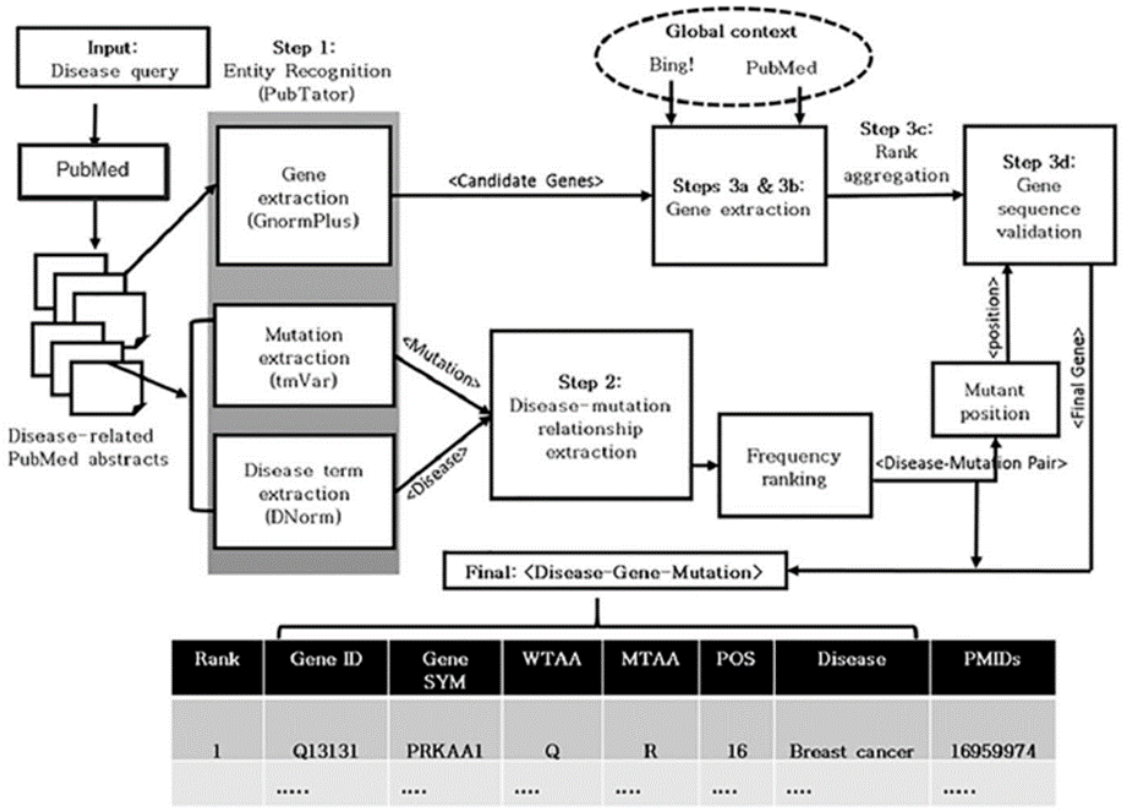

FIGURE 3: APPROCHE PROPOSÉE PAR SINGHAL, SIMMONS ET LU (2016)

Le travail de Singhal et al. (2016) montre une approche (décrite dans la Figure 3) similaire au système DiMeX avec quelques différences. La première est que l'outil développé n'utilise pas seulement du contenu des bases de données, mais également du contenu venant d'internet avec le moteur de recherche Bing ! (appelé « contexte global »). La deuxième est qu'une validation des séquences de gènes est utilisée afin d'augmenter la confiance en la relation entre gène et mutation. Un autre point intéressant dans ce travail est la validation du système par comparaison avec un autre système semi-automatique, EMU par Doughty et al. (2011). Les F-scores calculés sont de 79,4% et 74% pour 2 jeux de données différents.

2.3.4 EXTRACTION D'ASSOCIATION PRODUIT CHIMIQUE/MALADIE

Dans le travail de Yifan, Chih-Hsuan et Zhiyong (2016), l'objectif est de rechercher des associations entre produits chimiques et maladies. Le meilleur F-score obtenu est de 61,01% en utilisant les ensembles d'apprentissage et de développement BioCreative V et une amélioration par ajout de données supplémentaires générées automatiquement. BioCreative V est un jeu de données mis à disposition par l'organisation BioCreAtIvE (Hirschman, Yeh, Blaschke, & Valencia, 2005) (Critical Assessment of Information Extraction systems in Biology). Cette organisation a pour but de participer à un effort à l'échelle de la communauté pour évaluer les systèmes d'extraction de texte et d'extraction de l'information appliqués au domaine de la biologie.

L'approche utilise cette fois-ci un « Support Vector Machine » (Cortes & Vapnik, 1995) pour construire un modèle de classification. Le but étant de faire un classeur qui juge si une paire donnée produit chimique-maladie est cohérente et de le mettre à disposition du grand public, les auteurs pensent que ce travail illustre l'utilisation d'annotations au niveau des documents conservés dans les bases de données biomédicales existantes, qui sont largement négligées dans le développement de systèmes de « text mining ».

2.3.5 EXTRACTION D'ÉVÈNEMENTS BIOMÉDICAUX

Un travail proche du problème de recherche posé dans ce mémoire est certainement le travail de Bui et Sloot (2012). En effet, la première chose que l'on peut noter est qu'au lieu de simplement extraire des relations protéine-protéine, drogue-drogue, le système conçu permet d'extraire ce que

les auteurs appellent des « évènements » biomédicaux. Un évènement est défini comme suit : chaque évènement consiste en un déclencheur, un type et un ou plusieurs arguments. Deux exemples simples sont donnés dans l'article, représentés en Figure 4.

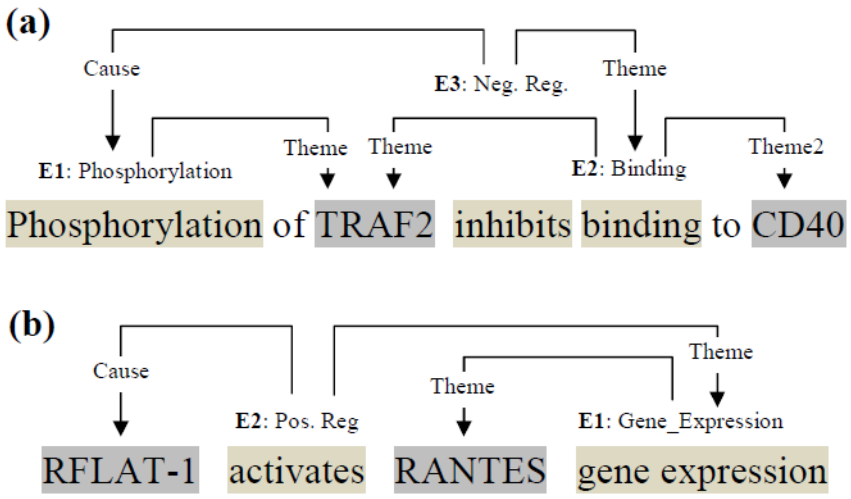

FIGURE 4: EXEMPLES DE 3 TYPES D'ÉVÈNEMENTS : SIMPLES (E1), LIAISON (E2) ET RÉGLEMENTAIRES (E3).

Dans la figure, (a) et (b) sont des exemples où les types d'arguments diffèrent : l'évènement E3 de (a) a deux évènements comme arguments alors que l'évènement E2 de (b) a une protéine et un évènement comme argument. Les résultats obtenus dans ce travail sont des F-score de 52,34% et de 53,34% pour deux différents jeux de données.

2.3.6 EXTRACTION DE SYMPTÔMES

Le travail se rapprochant le plus de notre problématique est certainement celui de Holat et al. (2016). De la même manière que dans ce présent mémoire, l'objectif est d'extraire des éléments cohérents et significatifs des publications et dans le cas de ce travail, cela concerne l'extraction et la reconnaissance de symptômes dans les publications en rapport avec les maladies rares. En utilisant 100 publications par maladie, sur 100 maladies sélectionnées par un expert, les auteurs sont parvenus à obtenir un F-score de 29,38%.

Un point intéressant est que ce travail utilise deux sources de données, dont nous parlerons plus tard à la section 4.3, OrphaData et Human Phenotype Ontology, qui représentent respectivement une source de données reconnue en ce qui concerne les maladies rares et un inventaire de symptômes. Ce travail diverge sur de nombreux points comparativement au nôtre. Premièrement, la reconnaissance de symptômes qui a été implémentée ne suit pas l'hypothèse H4 que nous avons posée. Pour rappel, cette hypothèse dit, de manière formelle : « Ce qui est extrait d'une publication en rapport avec une maladie M est lié à la maladie M ». Concrètement, cela veut dire que si on extrait un symptôme S d'une publication en rapport avec la maladie B, S est considéré comme un symptôme de la maladie B. En ajoutant à cela le fait que ce travail utilise uniquement les résumés de publications et qu'il ne considère que 100 maladies rares, il sera quasiment impossible de comparer leurs résultats avec les nôtres. Les méthodes utilisées sont le « pattern mining » (Aggarwal & Han, 2014) ainsi que le « sequence labelling » (Erdogan, 2010), ces approches sont comparées et combinées pour obtenir le meilleur résultat possible.

Un dernier travail, celui de Martin, Battistelli et Charnois (2014), a également tenté d'extraire les symptômes des publications. Un F-score de 36,8% est atteint sur 25 résumés pris au hasard en utilisant des méthodes de « pattern mining » et de « natural language processing » (Chowdhury, 2003). En plus d'expliquer et de tester leur approche, les auteurs mettent l'accent sur la difficulté de définir précisément ce qu'est un symptôme. D'abord, il n'y a pas de différence morphologique ou syntaxique entre un signe d'une maladie et un symptôme d'une maladie (un symptôme est remarqué et décrit par un patient), la différence n'est que sémantique. Ensuite, les symptômes prennent des formes linguistiques extrêmement variées. Martin et al. (2014) précisent que dans sa forme la plus simple, un symptôme est un nom, qui peut être complété par des compléments, tels que des adjectifs ou d'autres noms. Ils peuvent également apparaître sous d'autres formes plus complexes, allant d'une seule expression à une phrase entière.

2.4 DISCUSSION

Le tableau suivant résume l'état de l'art, selon les différents critères d'analyse :

TABLEAU 1: COMPARAISON DES TRAVAUX

Auteur	Modèle	Connaissances extraites	Performance	Utilisation du texte des publications
(Cha et al., 2016)	Collection de données + prétraitement + apprentissage par 3 méthodes + classification + évaluation	Résultats supplémentaires de recherche sur PubMed	+ 40% sur résultats de recherche	Partielle (titre, PMID, et résumé)
(Li et al., 2016)	Collection de données + prétraitement + apprentissage par 4 méthodes + classification + évaluation	Associations entre protéines	79,3%	Partielle (titre, PMID, et résumé)
(Mahmood et al., 2016)	Collection de données + « *natural language processing* » + évaluation	Associations mutations-gène-maladie	88%, 91% et 89%.	Partielle (titre, PMID, et résumé)
(Singhal et al., 2016)	Collection de données + prétraitement + « *machine learning* » + évaluation	Associations mutations-gène-maladie	79,4% et 74%	Partielle (titre, PMID, et résumé)
(Yifan et al., 2016)	Collection de données + prétraitement + apprentissage par SVM + classification + évaluation	Associations produit chimique-maladie	61,01%	Partielle (titre, PMID, et résumé)
(Bui & Sloot, 2012)	Collection de données + prétraitement + apprentissage de règle de sens+ combinaison des règles + évaluation	Faits (Évènements biomédicaux)	52,34% et 53,34%	Complète et partielle
(Holat et al., 2016)	Collection de données + projection avec dictionnaire + apprentissage avec CRF et « pattern mining » + combinaison des approches + évaluation	Symptômes de maladies rares	29,38%	Partielle (titre et résumé)
(Martin et al., 2014)	Collection de données + projection avec dictionnaire + apprentissage avec « pattern mining » + évaluation	Symptômes de maladies rares	36,8%	Partielle (titre et résumé)

D'après le Tableau 1, si l'on considère le critère de l'utilisation du texte des publications, on remarque que seuls les travaux de Bui et Sloot (2012) utilise le texte complet des publications. Il pourrait donc être intéressant de réutiliser la méthodologie employée et de voir comment les obstacles liés à la grande quantité d'informations non pertinentes sont surmontés.

Un deuxième point à noter est qu'en comparant les travaux de Mahmood et al. (2016) et de Singhal et al. (2016), on voit qu'au niveau des performances, c'est le travail de Mahmood et al. (2016) qui l'emporte de 10% environ, ce qui n'est pas négligeable. Cette affirmation pourrait alors laisser penser, au vu des différences entre les 2 travaux, que l'approche « natural language processing » fonctionne mieux que l'approche « machine learning ». Cette nouvelle affirmation pourrait également être intéressante à vérifier dans le cadre du projet de recherche, car elle permettrait de savoir quelle est la meilleure approche à adopter.

Plus globalement, on peut voir que les modèles utilisés par tous les travaux sont proches, les seules différences notables concernent le nombre de méthodes d'apprentissage utilisées et le type d'apprentissage utilisé.

Ensuite, si l'on considère les connaissances extraites, les travaux de Bui et Sloot (2012) sont une tentative de généralisation des autres travaux, car un évènement biomédical pourrait englober des associations d'éléments biomédicaux (gènes, protéines, maladies, mutations). Étant donné que ces travaux ont bientôt 6 ans et que les performances restent moyennes, il serait tout aussi intéressant de continuer dans cette voie afin de couvrir le plus de faits biomédicaux pertinents que possible.

Pour continuer, le travail de Cha et al. (2016) est à part, car il permet de son côté une extraction de plus de résultats de recherche sur PubMed. La problématique d'augmenter le nombre de résultats de recherche de PubMed pourrait être utile pour bénéficier d'un nombre de résultats significatif (plus de 50 comme dit précédemment) pour une maladie rare donnée.

Enfin, dans les travaux de Holat et al. (2016) et Martin et al. (2014), on constate que les performances ne sont pas très élevées (toujours inférieur à 40%). Augmenter les performances de l'extraction de symptômes est donc une problématique qui reste à creuser. D'ailleurs, en plus de concerner les maladies rares (dont il est difficile de trouver de la connaissance), les symptômes sont

des entités qu'un public non-expert peut facilement comprendre. C'est ce genre d'éléments que notre projet de recherche souhaite extraire dans les publications scientifiques.

2.5 CONCLUSION

Dans ce chapitre, tous les travaux présentés nous ont permis de savoir quels sont les différents types d'éléments qui peuvent être extraits des écrits scientifiques. Les éléments extraits et les approches, qui sont toutes différentes, montrent que des avancées scientifiques ont été faites dans l'extraction d'éléments des publications biomédicales. Cependant, on voit clairement qu'un élément reste difficile à extraire : les symptômes. Très peu de travaux traitent de leur extraction, et nous pensons que cela représente une grande opportunité de recherche. Une autre remarque importante est que la quasi-totalité des travaux présentés utilise partiellement les publications, en ne prenant que le titre et le résumé. Nous croyons qu'utiliser le texte complet permettrait d'obtenir de meilleurs résultats.

Le travail que nous proposons traitera donc de l'extraction de symptômes dans le texte complet des publications liées aux maladies rares. L'approche que nous envisagerons pour l'extraction n'est pas limitée aux symptômes, mais elle représente une première tentative pour ce travail de recherche.

CHAPITRE 3

LE « TEXT MINING »

Nous avons vu au chapitre 2 qu'il existe peu d'approches entourant l'extraction automatique de symptômes associés aux maladies rares. Cette extraction fait partie d'un grand domaine, le « text mining » ou la fouille de texte. Avant d'entrer dans le vif du sujet de ce mémoire, il apparaît important de donner les bases de cette discipline ainsi que les notions essentielles à la bonne compréhension des contributions apportées. Dans ce chapitre, nous commencerons par définir le « text mining », les différentes tâches que l'on peut réaliser ainsi que les concepts à connaître pour le faire correctement.

3.1 DÉFINITIONS

Dans cette section, le « text mining » et différentes disciplines pertinentes à ce sujet de recherche sont présentées. Ces définitions visent à permettre au lecteur de se faire une idée des possibilités, des buts et des enjeux que le « text mining » implique. Les détails théoriques ont volontairement été omis dans cette section.

3.1.1 « TEXT MINING »

Le « text mining » est un ensemble de méthodes, de techniques et d'outils développés pour exploiter les documents non structurés (Sailaja, Padmasree, & Mangathayaru, 2016) que sont les textes écrits que ce soit des documents Word, des courriels, des documents de présentation de type PowerPoint, des rapports ou encore des publications scientifiques. Ces documents, facilement compréhensible pour les humains, restent pourtant difficiles à comprendre par les machines. (Allahyari et al., 2017)

Afin d'extraire de l'information de ces documents non structurés, le « text mining » s'appuie sur des techniques d'analyse du langage naturel que nous définissons un peu plus loin (« natural language processing » ou NLP). Le « text mining » peut avoir plusieurs objectifs : extraction d'informations, suivi de sujet, synthèse automatique, catégorisation, faire des liens entre des concepts,

segmentation, visualisation d'information, réponse à question. La Figure 5 illustre le fonctionnement général du « text mining » :

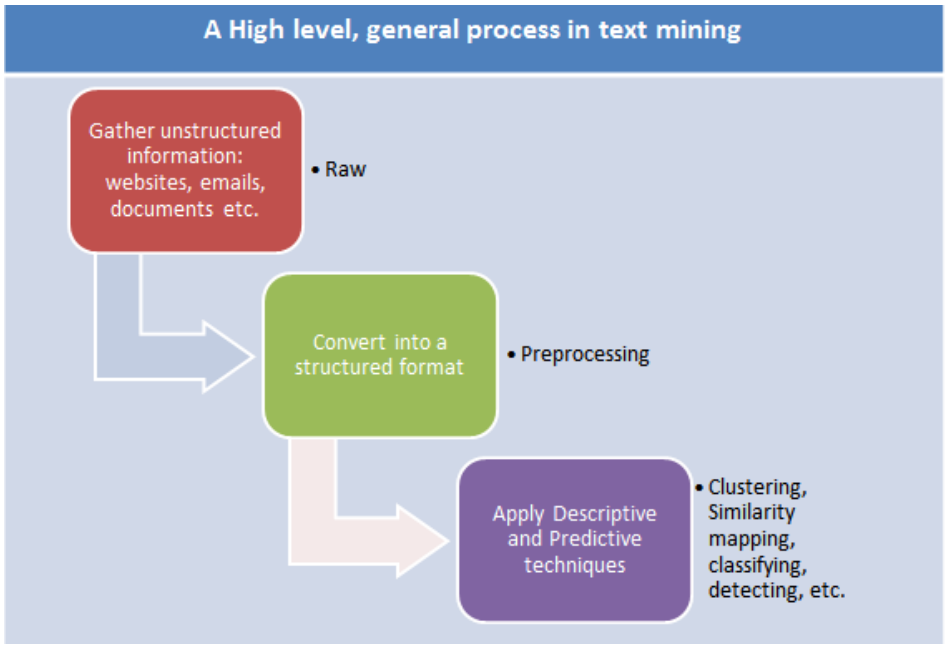

FIGURE 5: PROCESSUS À HAUT NIVEAU DU TEXT MINING (DESHPANDE, 2012)

En résumé, la Figure 5 précise un processus qui part de données non structurées (le texte), pour les convertir en données structurées grâce à une phase de prétraitement. Une fois le texte converti en un format structuré (aussi appelé modèle de représentation), on peut appliquer diverses techniques d'extraction de connaissances telles que des techniques d'apprentissage (classification, segmentation, règles d'association, etc.).

3.1.2 « TEXT MINING » DANS LE DOMAINE BIOMÉDICAL

Le « text mining » n'est pas étranger au domaine biomédical. En effet, depuis plusieurs années, beaucoup de projets ont vu le jour pour aider la communauté scientifique de la médecine et du biomédical en général. On peut par exemple citer Bui et Sloot (2012), dont nous avons déjà parlé dans notre revue de littérature, qui parviennent à extraire des publications des éléments complexes appelés évènements biomédicaux, qu'on peut interpréter comme des faits marquants et importants dans le domaine biomédical. (Exemple : « La protéine A inhibe l'expression du gène B »).

Comme mentionné dans l'enquête réalisée par Cohen et Hersh (2005), l'objectif du « text mining » dans les écrits biomédicaux est de permettre aux chercheurs d'identifier plus efficacement les informations nécessaires, de découvrir les relations cachées par la grande quantité d'informations disponibles, et plus généralement de faire passer le fardeau de la surcharge d'information du chercheur à l'ordinateur en appliquant des méthodes algorithmiques, statistiques et de gestion des données à la vaste quantité de connaissances biomédicales qui existe dans la littérature ainsi que dans les bases de données biomédicales.

3.1.3 « NATURAL LANGUAGE PROCESSING »

Un sujet intimement lié au « text mining » est celui du « natural language processing » (NLP). Pour être plus spécifique, le « text mining » (Kao & Poteet, 2007) pourra utiliser des méthodes de NLP, car c'est une manière d'envisager le traitement du texte par le rapprochement avec le langage humain. Pour être plus formel, le NLP est une méthode de traduction entre le langage « informatique » et humain. Il s'agit de la discipline d'intelligence artificielle qui permet à l'ordinateur de comprendre les concepts et le sens des textes écrits. En d'autres termes, le NLP automatise le processus de traduction entre les ordinateurs et les humains. Le NLP est souvent décrit comme un processus de plusieurs étapes, et d'après Yvon (2016), voici ces différentes étapes :

- Segmentation du texte en unités lexicales (en mots par exemple)
- Traitement lexical : identifier les composants lexicaux, et leurs propriétés
- Traitement syntaxique : identifier des constituants (groupe) de plus haut niveau, et les relations (de dominance) qu'ils entretiennent entre eux
- Traitement sémantique : construire une représentation du sens de cet énoncé, en associant à chaque concept évoqué un objet ou une action dans un monde de référence (réel ou imaginaire)
- Traitement pragmatique : identifier enfin la fonction de l'énoncé dans le contexte particulier de la situation dans lequel il a été produit

3.1.4 « NAMED ENTITY RECOGNITION »

L'application qui nous intéresse le plus pour ce mémoire se nomme la « named entity recognition ». La « named entity recognition » ou NER est une méthode permettant d'identifier dans un texte tous les mots ou groupes de mots qui appartiennent à la même catégorie. Les catégories peuvent être de toutes les natures imaginables : symptômes, noms de personnes, noms d'organisation, lieux, distances, gènes, dates, animaux, etc. Elle a, par exemple, été utilisée dans les travaux de Tanabe et Wilbur (2002), afin d'identifier les gènes et les protéines dans les publications biomédicales. La

Une des applications de la NER au-delà du fait d'identifier des termes appartenant à des catégories est la pseudonymisation. La pseudonymisation est l'art d'empêcher l'attribution de données à une personne physique sans avoir recours à des informations supplémentaires (Cadiet, 2017). Ainsi, en identifiant tout ce qui est considéré comme personnel et en le masquant, on peut permettre la libre circulation de documents, qui, initialement, contenaient trop d'informations sur la vie privée de certaines personnes.

La NER peut se faire de manières très variées, que ce soit par la création de règles d'identification faite à la main ou par des approches de « machine learning ». Les travaux de Nadeau et Sekine (2007) font état de ces différentes approches.

3.1.5 « TEXT CLASSIFICATION »

Une autre tâche du « text mining » est la classification de texte ou « text classification ». Elle consiste en la classification de tout ou partie d'un texte pour savoir si le document présente certaines caractéristiques. En général, elle est utilisée pour savoir si le texte parle d'un certain sujet ou si une certaine information est présente. Par exemple, lors d'une compétition faisant partie du « Knowledge Discovery in Databases » (KDD) Challenge Cup 2002, on a cherché à savoir, grâce à un algorithme de « text mining », s'il y a présence ou non de preuves expérimentales de produits de gènes de drosophiles dans le document. Lors de cette compétition, Yeh, Hirschman et Morgan (2003) sont parvenue à obtenir un F-score de 78 %.

3.1.6 SYNONYMES ET EXTRACTION D'ABRÉVIATIONS

La littérature biomédicale ne cesse de grandir et avec celle-ci, la croissance de sa terminologie. Il peut donc être intéressant de pouvoir réunir tous les termes ayant le même sens, car cela pourrait permettre d'obtenir un gain d'efficacité certain pour faire des similitudes et des rapprochements entre différentes publications. De nombreux travaux abordant cette thématique ont été développés. Ces travaux utilisent par exemple le principe de proximité des phrases/abréviations, des règles de correspondance de patterns manuelles ou encore de modèles avec caractéristiques spécifiques d'abréviation comme décrit par Cohen et Hersh (2005).

3.2 CONCEPTS

Maintenant que la terminologie et les définitions entourant le « text mining » ont été données, nous pouvons entreprendre l'approfondissement du processus complet du « text mining ». Plus spécifiquement, nous décrirons les éléments constituant les grandes étapes du processus de « text mining » illustrés dans la Figure 5, à savoir le prétraitement et les algorithmes existants.

3.2.1 PRÉTRAITEMENT

Le prétraitement est une étape importante du « text mining ». Son rôle est de préparer le texte brut afin de le transformer en données structurées dans l'optique de pouvoir être utilisés par d'autres algorithmes plus tard. Il existe de nombreuses manières de faire le prétraitement ; cette section a pour but de les présenter.

3.2.1.1 TOKENISATION

La tokenisation est l'acte de décomposer une séquence de chaînes de caractères en morceaux tels que mots, mots-clés, phrases, symboles et autres éléments appelés « token ». Les « tokens » peuvent être des mots individuels, des expressions ou même des phrases entières. Les « tokens » deviennent en général l'entrée d'un algorithme ou d'une autre phase de prétraitement. Par exemple, si je dispose de la phrase suivante : **« Charles rédige son mémoire »**. Ce qui sera retourné par une tokenisation en mots sera une liste de mots que voici : **[« Charles », « rédige », « son », « mémoire »]**. La tokenisation est principalement utilisée dans l'identification des mots ou des

phrases significatifs dans un texte. La définition de ce qui est significatif ou non, quant à elle, sera de la responsabilité du prochain processus qui traitera les « tokens ».

3.2.1.2 SUPPRESSION DES « STOP WORDS »

Un grand nombre de mots utilisés de manière fréquente dans les documents sont très peu significatifs ; souvent parce qu'ils permettent seulement de faire des liaisons entre mots. On appelle ces mots des « stop words ». Voici quelques exemples de « stop words » dans la langue française **: [« de », « des », « ou », « et »]** et d'autres en anglais **[« a », « an », « any », « all », « am », « out », « the »]**. Dû à leur fréquence d'apparition élevée dans les textes, ils peuvent rendre l'analyse de texte et le « text mining » plus difficile à mettre en place. Une solution est alors de les supprimer purement et simplement du texte à traiter à partir d'une liste préétablie, comme celle établie par Fox (1989).

Il faut savoir qu'une suppression des « stop words » est souvent précédée d'une tokenisation. En effet, pour supprimer des mots d'un texte, il faut avoir déjà identifié chaque mot, ce que la tokenisation permet de faire. Cependant, il est important de noter que le développement d'une liste des « stop words » pour une langue peut s'avérer difficile ou incompatible en fonction des besoins des prochains traitements à effectuer. Par exemple, l'article défini « de », qui pourrait s'avérer utile dans le cas d'une analyse sémantique (pour une relation d'appartenance par exemple), ne l'est pas du tout dans le cadre de la classification de documents. En effet, en classification de document, on utilise bien plus souvent les noms, car ils permettent de saisir le texte globalement.

3.2.1.3 « STEMMING »

Une autre méthode de prétraitement fréquemment employée est celle que l'on appelle le « stemming ». Le « stemming » est un procédé qui consiste à réunir les mots d'une même famille en une même représentation qui est la racine de ces mots. Par exemple, les mots **« vivre », « vivant », « vivable »** pourraient tous être réduits à la même racine **« viv »**. Elle conduit à des formes qui ne sont pas des mots, c'est donc un traitement final, qui empêche de faire un traitement approfondi des termes du texte.

Il existe plusieurs types d'algorithmes pour faire du « stemming », c'est-à-dire plusieurs types de « stemmer ». L'un des premiers est celui créé par Lovins (1968). Dans ce papier, la base du fonctionnement du « stemming » a été établie et a permis le développement d'autres algorithmes

plus performants, notamment avec la création de Snowball, qui est un langage de programmation créé par Porter (2001) spécialement dédié à l'implémentation d'algorithmes de « stemming ».

Le procédé est très utilisé par les moteurs de recherche pour améliorer la recherche d'informations. Les mots-clés d'une requête sont remplacés par leur racine. On peut ainsi regrouper plusieurs variantes d'un terme en une unique forme de représentation, ce qui va réduire la taille du dictionnaire utilisé.

3.2.1.4 LEMMATISATION

La lemmatisation est une opération très proche du « stemming », dans le sens où elle réduit un ensemble de mots d'une même famille en une forme unique. La différence principale est qu'au lieu de prendre la racine du mot, on va opter pour sa forme canonique. En français, passer à la forme canonique revient à :

- Passer les verbes à l'infinitif
- Passer les autres mots au masculin singulier

Cela revient à supprimer les notions de genre (masculin ou féminin), de nombre (singulier ou pluriel), de temps (présent, passé…) et de mode (indicatif, impératif…) des mots pour arriver à une forme unique. De cette manière, par lemmatisation, les mots **[« a », « aurait », « eussions eu », « avait »]** sont réduits au mot **« avoir »**.

3.2.1.5 ENCODAGE POUR LA « NAMED ENTITY RECOGNITION »

Dans des cas spécifiques comme la NER, il peut être nécessaire d'annoter le texte pour former un jeu d'apprentissage. Ce jeu d'apprentissage va servir à entraîner des algorithmes de classification en leur donnant la correspondance entre chaque « token » rencontré avec une étiquette. L'important pour ce jeu d'apprentissage, finalement, c'est que les tags permettent de signifier les termes appartenant à la catégorie voulue et les termes qui n'y appartiennent pas. Par exemple, si notre jeu d'apprentissage contient la phrase **« Charles a un rhume »** et que le but est de reconnaître des symptômes, le mot **« rhume »** doit être mis en valeur, car c'est effectivement un symptôme.

Ce jeu d'apprentissage est généralement encodé de trois façons différentes : IO, BIO et BMEWO. L'encodage choisi aura un impact sur la qualité de l'annotation des termes ainsi que sur la lourdeur du traitement qui suivra l'encodage. En général, un encodage qui donne plus d'informations,

et qui a donc nécessité plus d'effort permettra d'alléger l'effort de calcul du prochain traitement qui sera effectué avec le texte encodé.

3.2.1.5.1 ENCODAGE IO (Breckbaldwin, 2009)

L'encodage le plus simple est l'encodage IO, qui marque chaque « token » comme étant (I_X) un type particulier de type d'entité nommée X ou dans aucune entité (O). Ce codage est défectueux, car sachant qu'une même entité peut être constituée de plusieurs mots, il ne peut pas représenter deux entités l'une à côté de l'autre, car il n'y a pas de délimitation possible. (Dans le texte : « chat chien oiseau », l'encodage IO ne permet pas de différencier les 3 animaux)

Par exemple, si le but est d'identifier les personnes dans du texte, avec la phrase « Jean-Charles et Mathieu vont rendre visite à Albert », on peut représenter l'encodage IO par le tableau suivant :

TABLEAU 2: EXEMPLE D'ENCODAGE IO

Jean	I_PERSONNE
-	I_PERSONNE
Charles	I_PERSONNE
et	O
Mathieu	I_PERSONNE
rendent	O
visite	O
à	O
Albert	I_PERSONNE

3.2.1.5.2 ENCODAGE BIO (Breckbaldwin, 2009)

L'encodage BIO reprend exactement l'encodage IO, mais en ajoutant un tag B_X au premier « token » signifiant que l'on se trouve au début de l'entité. Grâce à cette légère modification, deux entités l'une à côté de l'autre peuvent être délimitées.

Avec le même exemple que précédemment, cela donnerait le tableau suivant :

TABLEAU 3: EXEMPLE D'ENCODAGE BIO

Jean	B_PERSONNE
-	I_PERSONNE
Charles	I_PERSONNE
et	O
Mathieu	B_PERSONNE

rendent	O
visite	O
à	O
Albert	B_PERSONNE

3.2.1.5.3 ENCODAGE BMEWO (Breckbaldwin, 2009)

Enfin, l'encodage BMEWO reprend les principes de BIO en ajoutant différentes notions. Ces dernières permettent d'alléger le traitement de l'encodage dans les cas où ce traitement nécessite un comportement particulier pour les entités à mot unique et aux termes appartenant au milieu et à la fin de l'entité. Plus concrètement, l'encodage BWEMO remplace le tag I_X par les tags suivants :

- E_X : Fin de l'entité (End of entity)
- M_X : Milieu de l'entité (Mid of entity)
- W_X : Entité à un seul « token »

En reprenant l'exemple précédent, cela donne :

TABLEAU 4: EXEMPLE D'ENCODAGE BMEWO

Jean	B_PERSONNE
-	M_PERSONNE
Charles	E_PERSONNE
et	O
Mathieu	W_PERSONNE
rendent	O
visite	O
à	O
Albert	W_PERSONNE

3.2.1.6 LE BESOIN DE PRÉTRAITEMENT POUR L'ANALYSE

Le prétraitement a trois principaux intérêts. D'abord, il permet de réduire la taille des données contenues dans les documents. On peut citer la suppression des « stop words » et le fait que le « stemming » réduit chaque mot en une chaîne de caractères plus petite. Ensuite, il permet d'améliorer l'efficacité des systèmes d'extraction d'information. Par exemple, les « stop words » sont inutiles au sens du texte, on peut trouver des mots similaires grâce au « stemming » et on peut créer un jeu d'apprentissage pour certains algorithmes. Enfin il peut permettre de créer des jeux d'apprentissage pour les algorithmes de classification par l'utilisation d'un encodage.

3.2.2 MODÈLES DE REPRÉSENTATION

Un second élément essentiel au « text mining » est celui des modèles de représentation du texte. Quand on fait du « text mining », un soin tout particulier doit être attribué à la sélection de ce dernier. En effet, le choix du modèle aura des conséquences sur la facilité d'utilisation du texte, sur les algorithmes utilisés et sur le type d'applications possibles de ce dernier. Dans cette section, nous verrons les principaux modèles utilisés dans la littérature en partant du plus simple, la chaîne de caractère, en passant par des plus élaborés comme « bag of words », pour finir par le plus complexe « word embedding ».

3.2.2.1 CHAÎNE DE CARACTÈRES

La plus simple représentation d'un texte est la simple chaîne de caractères. Ce format a pour avantage d'être très facile à utiliser, c'est la forme initiale de tout texte en informatique. Cependant, sans transformations supplémentaires, il est difficile d'en tirer beaucoup d'informations. Il peut toutefois être utilisé dans certains cas, par exemple quand on dispose d'expressions régulières pour fouiller le texte (nous en parlerons en 3.2.3.1.1).

3.2.2.2 ENSEMBLE DE PHRASES

À la suite d'une « tokenisation » d'un texte en phrases, on obtient un ensemble de phrases ordonnées. Par exemple le texte : **« Charles rédige son mémoire. Il veut écrire 2 pages par jour. »** devient : **[« Charles rédige son mémoire », « Il veut écrire 2 pages par jour »]**. Ce format reste très facile à utiliser et permet de traiter chaque phrase séparément. Ce modèle est fréquemment utilisé comme premier modèle envisagé dans de nombreux travaux, car chaque phrase d'un texte représentant une partie du sens global du texte, il est intéressant de pouvoir les analyser indépendamment.

3.2.2.3 ENSEMBLE DE MOTS

De la même manière qu'en 3.2.2.2, on obtient une liste de mots ordonnés exploitable. Chaque mot participant au sens d'une phrase, il est également intéressant de pouvoir les analyser indépendamment. Ce modèle est souvent employé en complémentarité avec les ensembles de

phrases pour découper le texte en une structure ordonnée à deux niveaux, le texte devient un ensemble de phrases, chaque phrase étant elle-même un ensemble de mots.

3.2.2.4 « BAG OF WORDS »

La représentation « bag of words » est un format qui se focalise sur la fréquence des mots dans un texte. Après avoir réalisé une tokenisation, il suffit de parcourir la phrase et, à chaque « token » rencontré, on lui associe sa fréquence dans le texte. Par exemple, si on applique « bag of words » au texte suivant : **« Charles part faire les courses dans une grande surface. Il aime généralement faire les courses le lundi. »**. On obtient, au format JSON : **{Charles : 1, part :1, faire : 2, les : 2, courses : 2, dans : 1, une : 1, grande : 1, surface : 1, Il : 1, aime : 1, généralement : 1, le : 1, lundi : 1}**. L'ordre des « token » n'a pas d'importance dans cette représentation. Par exemple, la représentation suivante est strictement identique à la précédente : **{lundi : 1, Charles : 1, grande : 1, part :1, faire : 2, aime : 1, les : 2, généralement : 1, courses : 2, dans : 1, une : 1, surface : 1, Il : 1, le : 1}**.

En pratique, le modèle « bag of words » permet de calculer différentes mesures pour caractériser le texte. La plus souvent, à chaque « token », on associe sa fréquence brute dans le texte. Le choix d'utiliser la fréquence brute n'est pas toujours un bon choix, car il existe des termes très fréquents comme « de », « et », « un » qui reviennent si souvent que leurs fréquences « écrasent » les autres. Pour remédier à cela, on peut décider de normaliser ce terme avec TF-IDF que nous verrons plus loin ou d'utiliser une fréquence binaire. (1 : le mot est présent dans le texte, 0 : le mot est absent du texte)

3.2.2.5 « WORD EMBEDDING » ET WORD2VEC

Presque toutes les représentations précédentes représentent les mots comme des entités uniques. Or, un mot seul a peu de sens si son contexte n'est pas précisé. Par exemple le mot **« souris »** possède un sens différent dans les phrases suivantes : « Le chat chasse désespérément une **souris** pour la dévorer », « J'ai acheté un clavier et une **souris** pour mon ordinateur ». Pour permettre de faire la distinction, un ordinateur peut :

- Apprendre tous les sens de tous les mots d'une langue et faire une analyse sémantique de la phrase pour décider quel est le sens le plus proche.

- Analyser le contexte dans lequel le mot est employé en regardant les voisins de ce mot.

La première solution est fastidieuse, car en plus de stocker toutes les définitions des mots existants, l'analyse sémantique par un ordinateur d'un texte n'est pas encore une tâche qui peut se faire de manière rapide et efficace. En ce qui concerne la seconde solution, regarder le voisinage du mot aurait pu nous aider déterminer le sens. Si l'on considère que les mots « chat » et « chasser » font partie du contexte du mot **« souris »** dans la première phrase, il est assez naturel de penser que la souris désignera l'animal et non l'objet.

Cependant, utiliser les mots du contexte tel quel pose un problème de dimensionnalité (on parle aussi de « fléau de dimensions ») originellement identifié par Bellman (2013). En effet, le nombre de mots existants dans un texte ou un ensemble de textes peut être de l'ordre des dizaines de milliers de mots différents et chacun d'entre eux peut avoir des formes variées (singulier/pluriel, masculin/féminin, conjugaisons). La présence d'un grand nombre de dimensions implique souvent de nécessiter un trop grand nombre d'observations pour obtenir une couverture équivalente.

Afin de pallier ce problème, il existe une solution consistant à représenter les mots sous la forme d'un vecteur de nombres réels dans un espace de dimension beaucoup plus petit. De cette façon, on peut estimer que des mots qui apparaissant dans un contexte similaire ont des vecteurs correspondants relativement proches (en calculant la distance entre ces vecteurs). C'est ce qu'on appelle le « word embedding » ou plongement de mots introduit par Vukotic, Claveau et Raymond (2015).

Créé par une équipe de chercheurs menés par Tomas Mikolov chez Google et expliqué dans les travaux de Goldberg et Levy (2014), Word2vec est un modèle prédictif efficace pour l'apprentissage des « word embeddings » à partir de texte brut. Ce modèle utilise des réseaux de neurones à 2 couches et permet d'obtenir une représentation pour les mots et de leur contexte de manière efficace.

3.2.3 ALGORITHMES

Une fois notre texte structuré à l'aide d'un ou plusieurs modèles de représentation, il faut appliquer le troisième élément fondamental du processus de « text mining », à savoir les algorithmes.

Les algorithmes utilisés dans le cadre du « text mining » peuvent servir de nombreux buts. Cette section va tenter de les présenter et d'expliquer leur fonctionnement.

3.2.3.1 « NAME ENTITY RECOGNITION »

Présentée en 3.1.4, la « Name entity recognition » (NER) permet d'identifier les mots ou groupes de mots appartenant à une même catégorie. Cette section présentera les principaux algorithmes qu'un système de NER peut utiliser ainsi que leurs avantages et inconvénients.

3.2.3.1.1 « RULE-BASED » NER

La « rule-based » NER établit un ensemble de règles grammaticales s'appliquant à un certain type d'entité. Ensuite, en parcourant le texte, si une portion de texte respecte l'une des règles établies, on dira que cette portion fait partie du type d'entité associé à cette règle. On utilise pour cela ce qu'on appelle des expressions régulières. Par exemple, si notre but est d'identifier les adresses courriel dans un texte, en prenant l'expression régulière suivante : ***(\w[-._\w]*\w@\w[-._\w]*\w\.\w{2,3})*** (par Neimke (2003)) et en l'appliquant au texte suivant : *« Charles aimerait que sa demande de CAQ soit acceptée. Pour cela, il a laissé son adresse charles.cousyn1@uqac.ca sur le site de l'immigration du Québec »*, on obtient alors la chaîne de caractère suivante : « *charles.cousyn1@uqac.ca* », ce qui correspond bien à ce que nous souhaitions obtenir. Chaque caractère ou groupe de caractères d'une expression régulière (« \w », « * », « @ », « {2,3} », « . », « - », « [», «] », …) a une signification précise qui aura un impact sur les termes qui seront identifiés par celle-ci. (Par exemple, « \w » correspond à tous les caractères alphanumériques)

L'avantage des expressions régulières est que, une fois l'expression trouvée, il suffit de l'appliquer par l'utilisation d'un moteur d'expression régulière (que tous les langages ou presque possèdent) pour obtenir les mots ou groupes de mots faisant partie de la catégorie voulue. Malheureusement, dans de nombreux cas, il est impossible de créer une expression régulière capable de détecter les entités voulues. Pour illustrer, il n'existe pas d'expression régulière capable de décrire tous les mots ou groupes de mots appartenant à la catégorie des **animaux**. Théoriquement, cela pourrait être possible en écrivant une expression régulière contenant tous les mots ou groupes de mots possibles appartenant à la catégorie des animaux, mais cela revient à l'approche par dictionnaire qui sera traitée dans la section suivante. L'explication est qu'une expression régulière s'occupe d'analyser les

différents caractères du texte, elle ne s'occupe ni de la sémantique ni de la fréquence des mots, elle regarde simplement si tel ou tel ensemble de caractères correspond aux règles établies dans l'expression régulière.

3.2.3.1.2 DICTIONNARY NER

La deuxième possibilité est d'utiliser une approche basée sur un dictionnaire. On part du principe que l'on dispose de la liste complète des entités possibles et on cherche simplement à extraire les mots ou groupes de mots du texte qui se trouvent dans cette liste. Cette approche a l'avantage d'être très simple, car une fois le dictionnaire trouvé ou établi, il suffit de vérifier les groupements de mots du texte et de trouver des correspondances. L'inconvénient principal avec ce genre de méthode est qu'elle est difficilement généralisable, dans le sens où le dictionnaire utilisé est rarement exhaustif. Cette approche ne peut détecter que ce qu'elle « connaît » déjà par le contenu du dictionnaire, il devient donc impossible de découvrir de nouvelles entités correspondantes à la catégorie recherchée par notre système de NER.

Il existe une variante permettant de généraliser très légèrement l'approche par dictionnaire. Elle se nomme « Approximate Dictionnary-based » NER et consiste à faire la même chose que l'approche classique, mais en ajoutant le principe suivant. Un mot A sera reconnu s'il est présent dans le dictionnaire ou bien si la distance entre le mot A et au moins un mot du dictionnaire est inférieure à un seuil fixé à l'avance. Il faut bien comprendre que la notion de « distance » introduite dans cette phrase est ce qu'on appelle en mathématique une métrique. Une distance souvent utilisée est la distance de Levenshtein ; elle est par exemple utilisée par Heeringa (2004) pour mesurer les différences de prononciation entre différents dialectes. Sa définition est le coût minimal pour transformer un mot M en P par les opérations élémentaires suivantes :

- Substitution d'un caractère de M par un caractère différent de P
- Insertion dans M d'un caractère de P
- Suppression d'un caractère de M

Par exemple si on prend les mots « NICHE » et « CHIENS », la transformation se fait en deux étapes :

- Suppression de N et I → CHE ;

- Insertion de I, N et S → CHIENS.

On a donc 2 suppressions et 3 insertions ce qui fait une distance de 5. Dans ce cas, si la distance maximale tolérée (le seuil) est de 3, alors, dans un dictionnaire contenant uniquement le mot « NICHE », le mot « CHIEN » n'est pas reconnu, car la distance avec les mots du dictionnaire est trop grande.

L'« Approximate Dictionnary-based » NER est capable de généraliser dans le sens où elle est capable de trouver des mots ou groupes de mots que le dictionnaire ne connaît pas mais qui restent proches des mots d'origine. Toute la difficulté d'utiliser une telle méthode est dans le réglage du seuil. En effet, si on a un seuil trop bas (0 par exemple), cela reviendra à une approche par dictionnaire classique et si on prend un seuil trop haut, on détectera beaucoup de mots qui ne font pas partie de la catégorie désirée (des faux positifs).

3.2.3.1.3 MODÈLES GRAPHIQUES

La troisième possibilité est d'utiliser des modèles graphiques qui seront entraînés avec des textes où chacune des entités nommées est déjà identifiée au préalable. On peut citer l'utilisation de « Hidden Markov Models » (HMM) ou de « Conditionnal Random Field » (CRF).

HMM est un modèle graphique permettant de modéliser un processus supposé markovien contenant des états cachés. Une chaîne de Markov est une manière de représenter ce processus. Expliqué et illustré par Rabiner et Juang (1986), nous allons tenter d'en expliquer le fonctionnement dans ce mémoire. Le principe du processus markovien est le suivant : « L'information utile pour la prédiction du futur est entièrement contenue dans l'état présent du processus et n'est pas dépendante des états antérieurs ».

Illustrons ce principe : prenons un système qui peut être dans deux états distincts, E et A, et supposons que nous sommes initialement dans l'état E. En se posant la question : « Vais-je rester dans l'état E ou non ? », selon le principe énoncé ci-dessus, il suffit de savoir qu'on se trouve dans l'état E pour prédire le prochain état. C'est tout à fait vrai, car si l'on regarde la Figure 6 représentant la chaîne de Markov de cet exemple, on voit que les flèches partant de l'état E représentent des probabilités de passage au prochain état. Dans notre cas, il y a une probabilité de 0.3 de rester dans

l'état E et 0.7 de passer à l'état A. Il en est de même pour l'état A qui a une probabilité de 0.6 de rester dans l'état A et 0.4 de passer à l'état E.

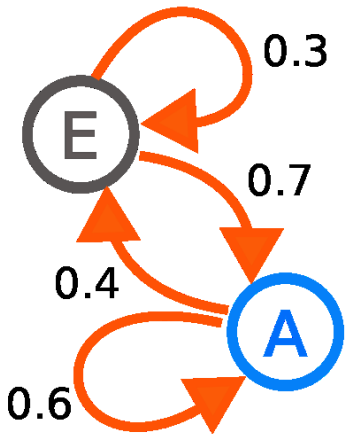

FIGURE 6: EXEMPLE DE CHAÎNE DE MARKOV

Ces probabilités permettant de passer d'un état à un autre se nomment probabilités transitionnelles. Il existe aussi souvent des probabilités permettant de savoir quel est l'état initial, on les nomme probabilités de départ. (Elles ne sont pas représentées dans la Figure 6, mais le sont dans la Figure 7). Dans le cas des modèles de Markov classiques, la connaissance de l'état présent permet de prédire l'état suivant ; ce n'est pas le cas des modèles de Markov cachés. La différence dans un HMM est que l'état présent n'est pas connu, à la place on a une variable d'observation plus ou moins corrélée avec les états cachés. Prenons le système suivant illustré par la Figure 7 :

- Deux états cachés possibles : la météo est **« Sunny »** ou **« Rainy »**
- La variable d'observation : l'activité du jour qui peut être **« Walk »**, **« Shop »** ou **« Clean »**

Comme dans un modèle de Markov classique, entre chaque état **« Sunny »** ou **« Rainy »**, il y a une probabilité de rester dans l'état actuel ou de changer d'état. Les probabilités en rouge et en bleu ont été ajoutées sur la représentation : elles représentent les chances que la variable d'observation ait la valeur j quand l'état caché est l'état i. Par exemple, quand on est dans l'état **« Sunny »**, il y a 60% de chance que l'activité du jour soit **« Walk »**. Ces probabilités sont appelées probabilités d'émission.

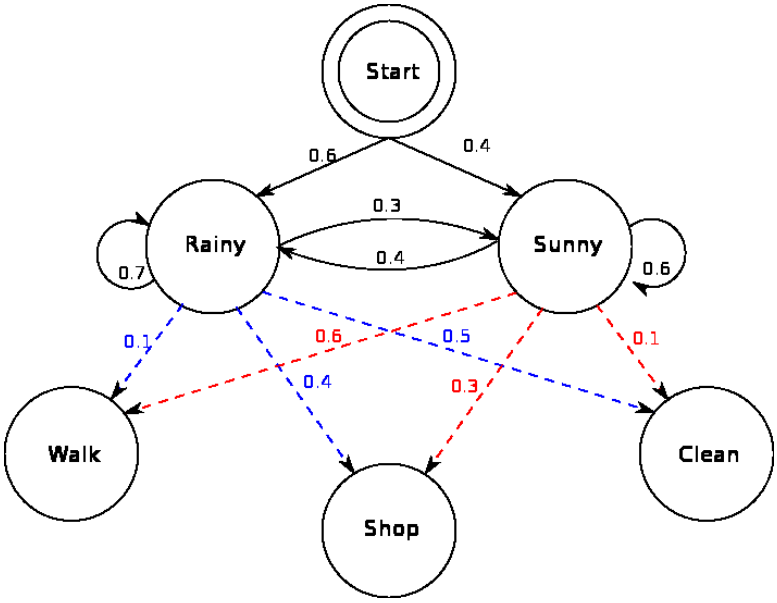

FIGURE 7: EXEMPLE DE CHAÎNE DE MARKOV CACHÉE

La question est maintenant de savoir en quoi les HMM peuvent être utilisés pour la tâche de NER. En fait, on part du principe de base qu'un processus markovien permet de traiter les séquences en général (des séquences d'états). Si on prend comme base qu'un texte est un ensemble de phrases et que chaque phrase est une séquence de mots, la tâche de NER avec un HMM revient à ce qui suit. Étant donné :

- Une phrase **O** contenant la séquence de mots suivante : [t_1,t_2,t_3,…,t_n]
- Un HMM entraîné pour reconnaître les entités appartenant à la catégorie **E**
- Les états pouvant être **E** ou **Autre**

On commence par déterminer la séquence d'états **S** la plus probable (appelée **S_max**) tel qu'illustré dans la Figure 8 :

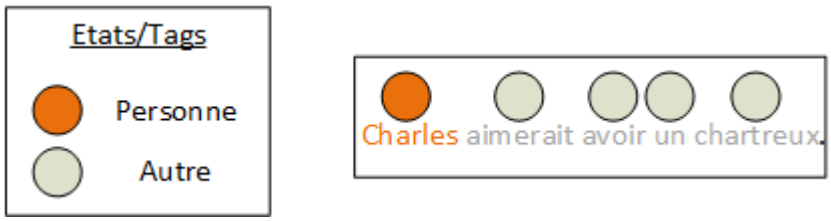

FIGURE 8: EXEMPLE DE NER AVEC UN HMM DONT LE BUT EST DE RECONNAÎTRE DES PERSONNES : LA SÉQUENCE DE COULEURS REPRÉSENTE LA SÉQUENCE DES ÉTATS ASSOCIÉS À LA SÉQUENCE DE MOTS DE LA PHRASE

Plus formellement, cela revient à trouver la séquence d'états **S_max** qui maximise la probabilité **P(S, O)**. Une fois la séquence trouvée (les ronds de couleur sur la Figure 8), tous les mots désignés comme appartenant à l'entité **E** recherchée sont extraits. L'étape la plus délicate est le fait de déterminer la séquence **S_max**. La première idée est bien évidemment de tester toutes les séquences possibles ([Autre, Autre, Autre, Autre, Autre], [Personne, Autre, Autre, Autre, Autre], [Personne, Personne, Autre, Autre, Autre],…). Le problème est que si la longueur de la séquence est longue, le nombre de combinaison à tester devient exponentiellement grand. Pour résoudre ce problème, il existe l'algorithme de Viterbi (Viterbi, 1967) qui permet de réduire énormément le nombre de combinaisons à tester (à le rendre linéaire). Ce qui précède est le principe de base de la NER avec les HMM. Le plus souvent, il y a également une étape d'étiquetage morphosyntaxique ou « Part-of-speech Tagging » (POS Tagging) (Voutilainen, 2003) qui a pour but d'améliorer les performances. Cela consiste en réalité à se passer des termes utilisés dans les phrases pour les remplacer un à un avec leur classe grammaticale (nom, verbe, article, préposition, adverbe, etc.).

D'autres modèles statistiques existent pour faire de la NER comme les « Conditionnal Random Field » (Lafferty, McCallum, & Pereira, 2001). La principale différence entre les HMM et les CRF est que ce sont des types de modèles différents. HMM est ce qu'on appelle un modèle génératif. C'est-à-dire : étant donné une variable observable **X** (mots) et une variable cible **Y** (Nom d'entité), un modèle génératif est un modèle statistique de la distribution de probabilité conjointe **X × Y**, notée **P(X, Y)**. Un CRF, quant à lui, est ce qu'on appelle un modèle discriminant. C'est un modèle statistique de de la distribution de la probabilité conditionnelle **Y sachant X**, notée **P(Y|X)**. Selon Sutton et McCallum (2012), l'avantage principal des modèles discriminants est qu'ils sont mieux adaptés à l'inclusion de caractéristiques riches et se chevauchant.

3.2.3.2 « TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY »

Nous avons mentionné précédemment qu'il est parfois utile de normaliser les fréquences de termes quand on utilise un modèle de représentation comme « bag of world » (en 3.2.2.4). Afin d'y parvenir, une des possibilités est d'utiliser la « Term Frequency – Inverse Document Frequency » ou son acronyme TF-IDF, initialement introduit par les travaux de Sparck Jones (1972), qui est une mesure statistique qui tente de refléter à quel point un mot/terme est important dans un corpus de texte. Instinctivement, il semble naturel de considérer que la fréquence d'un mot/terme dans un texte reflète l'importance dudit mot/terme dans le texte, il suffit d'avoir sa fréquence dans les textes, le nombre d'occurrences. De cette manière, dans les textes suivants : *« Le **chat** possède des sens très développés. Le **chat** est l'un des animaux dont l'ouïe est la plus fine.* », « Le **chat** est d'une nature très indépendante. Le **chat** domestique est une sous-espèce du chat sauvage (Felis silvestris). », on peut noter que l'un des mots qui ont la fréquence la plus élevée est le mot **« chat »**. En suivant le principe précédent, qui est, si on le précise : « Plus un mot est fréquent dans un texte, plus il est important pour le texte », on en déduit que le mot chat est important dans ce texte et qu'on a de bonnes raisons de penser que le texte a pour sujet une entité nommée **« chat »**. Cependant, on peut également noter que parmi les mots les plus fréquents, nous avons le mot « est ». Cependant, le terme « est » ne reflète rien d'important dans le texte, c'est un mot employé très souvent pour exprimer une idée, mais il ne représente pas l'idée.

TF-IDF tente de résoudre ce problème en pondérant la fréquence des mots par la fréquence inverse dans les documents. Puisque le terme « est » est si fréquent, sa simple fréquence risque de le mettre trop fortement en valeur alors que les termes qui ont plus de sens comme **« chat »** n'auront pas plus de poids. La fréquence inverse dans les documents peut être définie comme suit : « facteur diminuant le poids des termes qui apparaissent très fréquemment dans l'ensemble de documents et augmentant le poids des termes qui apparaissent rarement. »

3.2.3.2.1 FORMULATION MATHÉMATIQUE

La fréquence la plus souvent utilisée dans le cadre de TF-IDF est le nombre brut d'occurrences du terme ***t*** dans le document ***d***. Voici sa formulation mathématique :

$$TF(t,d) = f_{t,d} \qquad (1)$$

où $f_{t,d}$ est de nombre brut d'occurrences du terme t dans le document d

En ce qui concerne la fréquence inverse des documents (IDF) pour un terme t et un ensemble de documents D, la formule la plus souvent utilisée est :

$$IDF(t,D) = \log\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right) \qquad (2)$$

où $|\{d \in D : t \in d\}|$ est le nombre de documents du corpus D dans lequel le terme t apparaît et $|D|$, le nombre total de documents dans le corpus D.

La mesure TF-IDF est simplement le produit des deux facteurs précédents :

$$TF - IDF(t,d,D) = TF(t,d) * IDF(t,D) \qquad (3)$$

Il est important de noter que la mesure est liée à un triplet (terme t, document d, corpus de documents D) et non pas à un simple terme t.

3.2.3.2.2 LES DIFFÉRENTES VARIANTES DE TF-IDF

Les formules utilisées pour la « Term Frequency » et l'« Inverse Document Frequency » ne sont pas fixes, il existe un certain nombre de variantes. Ces variantes ont pour but d'avoir des valeurs positives et/ou d'amoindrir ou d'augmenter l'effet de ces facteurs. Les variantes de la « Term Frequency » sont représentées dans le Tableau 5 et sont décrites par les critères suivants :

- (1) : Normalisation entre 0 et 1
- (2) : Linéaire
- (3) : Amoindrit les effets des termes à haute fréquence (dans un document) comparativement aux autres
- (4) : La valeur du compte brut a un impact sur la mesure

TABLEAU 5: VARIANTES DE LA « TERM FREQUENCY »

	Formulation	(1)	(2)	(3)	(4)
Binaire	$\{0, 1\}$	OUI	NON	OUI	NON
Compte brut	$f_{t,d}$	NON	OUI	NON	OUI
Fréquence	$f_{t,d} / \sum_{t' \in d} f_{t',d}$	NON	OUI	NON	OUI
Normalisation logarithmique	$log(1 + f_{t,d})$	NON	NON	OUI*	OUI

Normalisation double 0.5	$\mathbf{0.5} + \mathbf{0.5} * \frac{\boldsymbol{f_{t,d}}}{\boldsymbol{max_{\{t' \in d\}} f_{t',d}}}$	OUI	OUI	OUI**	OUI
Normalisation double K***	$\boldsymbol{K} + (\mathbf{1} - \boldsymbol{K}) * \frac{\boldsymbol{f_{t,d}}}{\boldsymbol{max_{\{t' \in d\}} f_{t',d}}}$	OUI	OUI	OUI**	OUI

* : Avec la normalisation logarithmique, l'écart entre les termes très fréquents et les autres est proportionnellement réduit. Cela signifie qu'entre un terme très fréquent et un autre, le premier se voit attribuer une valeur plus amoindrie que le second. (Dérivée qui tend vers 0)

** : L'écart entre les termes très fréquents et les autres est réduit, mais pas proportionnellement. Cela signifie qu'entre un terme très fréquent et un autre, la variation de la valeur attribuée reste la même. (Dérivée constante)

*** : K est ce qu'on appelle un coefficient de lissage, c'est un nombre réel entre 0 et 1 dont le but est d'amortir la contribution du second terme

En ce qui concerne les variantes de l'« Inverse Document Frequency », elles sont représentées dans le Tableau 6 et sont décrites par les critères suivants :

- (1) : Toujours positif
- (2) : Amoindrit les effets des termes à haute fréquence (dans tout un corpus de documents) comparativement aux autres

TABLEAU 6: VARIANTES DE LA « INVERSE DOCUMENT FREQUENCY »

	Formulation	(1)	(2)
Unaire	$\mathbf{1}$	OUI	OUI
IDF	$\boldsymbol{log}\left(\frac{\boldsymbol{N}}{\boldsymbol{n_t}}\right)$	NON	OUI
IDF Smooth	$\boldsymbol{log}\left(\mathbf{1} + \frac{\boldsymbol{N}}{\boldsymbol{n_t}}\right)$	OUI	OUI
IDF Max	$\boldsymbol{log}\left(\frac{\boldsymbol{max_{\{t' \in d\}} n_{t'}}}{\mathbf{1} + \boldsymbol{n_t}}\right)$	NON	OUI
Probabilistic IDF	$\boldsymbol{log}\left(\frac{\boldsymbol{N} - \boldsymbol{n_t}}{\boldsymbol{n_t}}\right)$	NON	OUI

$\boldsymbol{N}$ désigne le nombre total de documents du corpus (strictement identique à $|\boldsymbol{D}|$) et $\boldsymbol{n_t}$ désigne le nombre de documents dans lequel le terme t apparaît (strictement identique à $|\{\boldsymbol{d} \in \boldsymbol{D} : \boldsymbol{t} \in \boldsymbol{d}\}|$ de la section 3.2.4.1).

3.3 CONCLUSION

En conclusion, dans ce chapitre, nous avons défini les notions importantes et nécessaires à la compréhension de ce travail de recherche. En particulier, nous avons présenté les différents pré-traitements qui peuvent être appliqués au texte que cela soit dans un but de réduction de taille des données, d'amélioration de l'efficacité des systèmes d'extraction d'information ou de préparation le texte au passage dans des algorithmes d'apprentissage. Nous avons ensuite vu les différentes manières de représenter le texte (modèles de représentation) allant du plus simple (chaîne de caractère) au plus élaboré (« word embeddings »). Enfin, nous avons présenté les différents algorithmes qui peuvent être appliqués en fonction du besoin, notamment en ce qui concerne la « named entity recognition » et la mesure statistique TF-IDF.

Parmi les approches vues pour chaque étape du processus de « text mining », certaines seront utilisées pour répondre à notre problématique et leur utilisation sera vue et discutée dans la prochaine section.

CHAPITRE 4

CONTRIBUTIONS DES TRAVAUX

Après avoir présenté le « text mining » dans son ensemble, ce chapitre a pour but d'expliquer quelle est la contribution du présent travail de recherche. Nous commencerons par parler de la problématique à traiter et des choix qui ont été faits pour ensuite parler des éléments extraits, des sources de données utilisées, des algorithmes et finalement des expérimentations.

4.1 PROBLÉMATIQUE

Tous les concepts présentés dans la section précédente ont été utiles dans la démarche de ce que ce mémoire pourrait apporter. Notre problématique est d'automatiser l'extraction de connaissances depuis les publications scientifiques sur les maladies rares dans le but d'obtenir de l'information fiable et pertinente.

Les connaissances pouvant être transmises à partir des publications scientifiques sur les maladies rares sont nombreuses. On peut extraire les dernières découvertes scientifiques qui peuvent être les gènes impliqués, les symptômes, l'espérance de vie, les médicaments utilisés. On peut également extraire des méta-informations sur ces publications, à savoir le nombre de publications mensuelles, les publications les plus citées, etc. Ce genre d'informations permet de connaître l'état de la recherche scientifique sur les maladies rares, ce qui est pertinent par rapport à leur rareté.

4.2 ÉLÉMENTS EXTRAITS

En posant cette problématique et avec la revue de littérature que nous avons effectuée, nous avons dû nous questionner sur le genre d'éléments que nous pourrons extraire des publications scientifiques liées aux maladies rares.

4.2.1 ANOMALIES PHÉNOTYPIQUES ET SYMPTÔMES

Le phénotype est l'ensemble des caractéristiques visibles d'un organisme : couleur des cheveux, des yeux, forme des oreilles ou du nez, taille, groupe sanguin, etc. On considère que le phé-

notype est en partie l'expression visible du génotype, définissant lui, le patrimoine génétique de l'organisme composé de différents gènes héréditaires. Toutefois, le phénotype peut aussi être fortement influencé par l'environnement : la peau peut brunir sous l'effet du soleil, un membre être perdu lors d'un accident.

Les anomalies phénotypiques sont des traits considérés « anormaux » développés chez un organisme. Leur présence dans la littérature scientifique fait partie de nos hypothèses de base, à savoir l'hypothèse H1 citée dans le chapitre 1. Une anomalie phénotypique se trouve être une notion relativement proche de ce qu'on appelle communément un symptôme et dans le cas des maladies rares, nous considérons ces notions comme similaires étant donné qu'elles sont, pour la grande majorité, d'origine génétique.

Aussi, pour la suite de ce mémoire, nous avons fait le choix d'extraire ces symptômes, mais d'utiliser le terme « anomalie phénotypique » pour être le plus exact possible.

4.2.2 MÉDICAMENTS

Les médicaments représentent toute substances ou compositions présentées comme possédant des propriétés curatives ou préventives à l'égard des maladies humaines ou animales. Dans le cas des maladies rares, chez les humains, on parle souvent de médicaments orphelins. Un médicament orphelin est un médicament censé traiter une maladie, mais la maladie est si rare que le médicament ne peut être développé dans un cycle de commercialisation classique.

Les médicaments orphelins sont un type d'élément que nous avons envisagé pour nos éléments à extraire des publications scientifiques. Cependant, pour des raisons de difficultés d'exploitation de jeux de données et de temps, nous avons décidé de laisser de côté ce genre d'éléments.

4.3 SOURCES DE DONNÉES UTILISÉES

Nous savons maintenant quel genre d'éléments vont être extraits et par conséquent, le système aura besoin de tout un ensemble de données qu'il faut trouver. Le chapitre qui suit va donc identifier les différentes sources de données qui ont été utilisées dans ce mémoire. Ces données servent différents buts, que cela soit pour obtenir la liste des maladies rares, les anomalies phénotypiques, les publications scientifiques et l'évaluation des résultats.

4.3.1 ORPHANET ET ORPHADATA

Dans la réalisation de ce projet de recherche, l'une des premières choses à faire était de s'occuper de la récupération d'informations sur les maladies rares. Nous avons opté pour l'utilisation d'Orphanet et d'OrphaData qui nous ont semblé contenir les données les plus complètes et à jour.

Orphanet est un portail de référence pour l'information sur les maladies rares et les médicaments orphelins, pour tous les publics. Il est dirigé par un consortium européen d'une quarantaine de pays, coordonné par une équipe française. Les équipes nationales sont responsables de la collecte d'informations sur les centres d'expertise, les laboratoires médicaux, les recherches en cours et les organisations de patients dans leur pays. Toutes les équipes d'Orphanet respectent la même charte de qualité. L'équipe de coordination française est responsable de l'infrastructure d'Orphanet, des outils de gestion, du contrôle qualité, de l'inventaire des maladies rares, des classifications et de la production de l'encyclopédie à des fins de consultations uniquement, l'ensemble des données sont accessibles depuis http://www.orpha.net.

La mission d'OrphaData, quant à lui, est de fournir à la communauté scientifique un ensemble de données complet, de haute qualité et librement accessible sur les maladies rares et les médicaments orphelins, dans un format réutilisable. Les données d'OrphaData sont une extraction partielle des données stockées dans Orphanet. En réalité, Orphanet fournit l'ensemble des données au public par son site web, mais OrphaData fournit gratuitement une partie des données dans des formats exploitables par les chercheurs (les données d'OrphaData peuvent être exploitées par un programme informatique). Ces données sont accessibles à l'adresse suivante http://www.orphadata.org/cgi-bin/index.php/.

Orphanet et OrphaData proposent différents jeux de données dont une partie est accessible gratuitement, à savoir :

- Un inventaire des maladies rares indexées avec OMIM, CIM-10, UMLS, MeSH, MedDRa
- Linéarisation des troubles
- Une classification des maladies rares établie par Orphanet, basée sur les classifications d'experts publiées

- Phénotypes associés aux maladies rares
- Maladies rares avec leurs gènes associés
- D'autres jeux de données sont également accessibles sur demande tels que les informations textuelles sur les maladies rares, une liste d'organisations de patients, une liste des médicaments orphelins et une liste des activités de recherche

Le format des jeux de données est dans la majorité des cas du XML ou du JSON. Il est important de noter que ces formats sont couramment utilisés pour la structure de donnée et permettent aux ordinateurs d'être programmés pour « comprendre » des informations. On dispose également d'une description de la structure de chacun des fichiers proposés à l'adresse http://www.orphadata.org/cgi-bin/docs/userguide.pdf. Grâce à ces deux éléments, nous avons réussi à exploiter sans difficulté les informations qu'Orphanet met à disposition.

4.3.2 JEU DE DONNÉES « PHÉNOTYPES ASSOCIÉS AUX MALADIES RARES » SUR ORPHADATA

Disponible à l'adresse http ://www.orphadata.org/cgi-bin/inc/product4.inc.php, le jeu de données « Phénotypes associés aux maladies rares » est un fichier au format XML mis à jour tous les mois qui, comme son nom l'indique, contient les phénotypes associés à chaque maladie rare. Il contient les phénotypes associés à près de 3300 maladies au moment de l'écriture de ce mémoire.

Ce jeu de donnée nous permet de connaître l'état actuel de la recherche en ce qui concerne les phénotypes. Il représente, ce qui sera pour nous la « vérité », les « véritables » phénotypes de nos maladies rares. En le comparant avec ce que nous sommes capables d'extraire, nous pourrons mesurer et nous représenter la qualité de notre extraction. Chaque phénotype est donné en utilisant les termes de la HPO (Human Phenotype Ontology), une terminologie normalisée et contrôlée couvrant les anomalies phénotypiques dans les maladies humaines (la section 4.3.5 décrira cette terminologie comme une autre de nos sources de données).

4.3.3 JEU DE DONNÉES « MALADIES RARES ET RÉFÉRENCES CROISÉES » SUR ORPHADATA

Le fichier, disponible à l'adresse http ://www.orphadata.org/cgi-bin/inc/product1.inc.php, représente la liste des maladies considérées comme rares de nos jours selon les critères d'Orphanet. Cette liste contient près de 7000 maladies au moment de l'écriture de ce mémoire.

Il est proposé au format XML ou JSON et est mis à jour tous les mois. Il représente la source à partir de laquelle nous allons chercher les maladies rares ainsi que les informations de bases à leur propos. Chaque maladie est identifiée par un identifiant nommé « OrphaNumber » ; ce sera l'identifiant qui sera utilisé dans notre base de données.

4.3.4 PUBLICATIONS SCIENTIFIQUES

Après avoir eu les informations de bases sur nos maladies rares, il nous manquait les publications scientifiques en lien avec ces dernières.

4.3.4.1 CHOIX ET JUSTIFICATIONS DE LA BASE DE DONNÉES

De nombreuses sources existent pour récupérer des publications scientifiques. Elles diffèrent principalement par leur contenu, le format du contenu et la méthode de récupération des données. Comme dit dans notre état de l'art, nous utiliserons seulement des publications dont le texte complet est disponible gratuitement, car nous croyons que cela pourrait améliorer les performances de l'extraction.

Le Tableau 7 compare différentes bases de données en fonction des critères suivants :

- Nombre de journaux : Représente le nombre de journaux que cette base de données met à disposition
- Texte complet : Précise si la base de données propose des publications avec le texte complet
- Format : Peut avoir 4 valeurs cumulables (**HTML Full Text** : Le texte est disponible en parcourant les éléments d'une page HTML, **PDF** : Le texte est disponible au for-

mat PDF, **XML** : Le texte est disponible en parcourant un fichier XML structuré, **Redirection** : La base de données redirige vers une autre base de données, le format est alors incertain)

- Moteur de recherche : Précise si un moteur de recherche est disponible pour chercher des publications scientifiques
- Présence d'une API : Précise si une interface de programmation est disponible (facilite grandement la récupération des publications)

TABLEAU 7: TABLEAU COMPARATIF DES BASES DE DONNÉES DE PUBLICATIONS SCIENTIFIQUES

Base de données	*Nombre de journaux*	*Texte complet*	*Format*	*Moteur de recherche*	*Présence d'une API*
Medknow	351	OUI	HTML Full Text	OUI	NON
PubMed Central (PMC)	3068	OUI	XML HTML Full Text PDF	OUI	OUI
Medscape	125	OUI	HTML Full Text	OUI	NON
Bioline	42	OUI	HTML Full Text	OUI	NON
Indmed	13	OUI	HTML Full Text	NON	NON
BMC (Biomed Central)	282	OUI	HTML Full Text PDF	OUI	NON
Geneva Foundation for Medical Education and Research	59	OUI	Redirection	OUI	NON
BMJ Journals	60	OUI	HTML Full Text PDF	OUI	NON
Elsevier Journals open archive	117	OUI	Redirection	NON	NON
LWW Journals	266	OUI	HTML Full Text PDF	OUI	NON

Parmi l'ensemble des critères cités ci-dessus, certains sont très discriminants. Premièrement, le nombre de journaux disponibles doit être le plus grand possible. Dans ce cas, c'est PMC qui l'emporte. Ensuite, concernant la présence du texte complet, chacune des bases comparées respecte cette condition. Le format a également son importance, car certains sont plus faciles à manipuler que d'autres. Le format XML est sans aucun doute le plus simple à utiliser pour un ordinateur,

car il permet d'organiser le contenu avec une structure en arbre. Le format HTML Full Text est également exploitable, car il reste organisé en structure d'arbre ; cependant il est plutôt destiné à l'affichage, sans oublier que le site web sur lequel il est disponible peut changer de structure HTML à tout moment, ce qui rendrait son exploitation sujette à des mises à jour. Enfin, le format PDF est le plus difficile à exploiter par un ordinateur, car l'organisation des sections, des titres, du texte varie souvent d'un journal à une autre, voire d'un article d'un même journal à un autre. Cela impliquerait également l'utilisation d'une librairie externe comme PDFlib TET créée par Merz (2018). Encore une fois, c'est PMC qui l'emporte sur ce critère.

La présence d'un moteur de recherche est capitale, car une base de données n'en possédant pas ne permet pas de trouver les articles en lien avec chaque maladie (à moins que la base de données ne contienne une section de publications pour chaque maladie, ce qui n'est pas le cas d'après nos recherches). Cela permet donc d'exclure Indmed et Elsevier Journals open archive de notre sélection.

Pour finir, il reste le critère de l'API. C'est une interface destinée aux programmeurs afin de leur mettre à disposition seulement les éléments et les actions dont ils ont besoin sur un système donné. Dans notre cas, le système est la base de données ; quant aux actions et aux éléments dont nous avons besoin, ils sont respectivement la récupération des publications et les publications en elles-mêmes. Une API facilite grandement l'étape de conception et l'implémentation de programmes informatiques de récupération de données. PMC est la seule base de données capable d'un tel service à condition de passer par l'interface « Entrez Programming Utilities » (E-utilities) créée par Sayers (2010). C'est par cette interface et en suivant le guide disponible en ligne que nous allons prélever les publications disponibles sur PMC.

On notera l'existence d'Europe PMC qui est une initiative du Europe P.M.C. Consortium (2015). Europe PMC regroupe les publications de PubMed et de PMC (PMC est une portion de PubMed avec des textes complets), qui propose également une API simple à utiliser. Le seul défaut de cette base qu'elle peut être moins complète que PMC dans le cas où les éditeurs ne donnent pas leur autorisation pour que leur contenu soit disponible dans cette base de publications.

4.3.4.2 PUBMED ET PMC

À l'origine, nous voulions utiliser les résultats de recherche du moteur de recherche PubMed. PubMed est un moteur de recherche permettant de naviguer dans la base de données MEDLINE, qui est la première base de données bibliographique de la « National Library of Medicine » (NLM) des États-Unis qui contient plus de 24 millions de références à des articles de revues en sciences de la vie axés sur la biomédecine. Le seul problème, sur PubMed, est que seulement une partie des publications scientifiques est disponible, à savoir le titre, les auteurs, les date et, le résumé.

PMC est une archive gratuite de la littérature des revues biomédicales et des sciences de la vie à la NLM des « National Institutes of Health » des États-Unis (US NLI). Tel que mentionné précédemment, PMC représente une portion de PubMed dont les publications sont disponibles avec le texte complet.

4.3.4.3 PARAMÉTRAGE ET HYPOTHÈSES

L'API E-utilities permet de préciser de nombreux paramètres quand on recherche des publications scientifiques. Parmi l'ensemble de ces paramètres, seul deux nous ont intéressés. Le premier paramètre qui nous intéresse est la limitation du nombre d'articles que l'on peut obtenir pour une maladie rare. Nous avons fait le choix de limiter ce nombre à 1000 publications par maladie pour des questions de stockage dans notre base de données et parce que cette quantité nous semble suffisante pour réaliser notre extraction. Ce paramètre se nomme retmax, et pour l'utiliser, il nous suffit d'ajouter **retmax=1000** dans nos requêtes.

Le deuxième paramètre qui nous intéresse est la méthode de tri des résultats. En effet, si on se limite à 1000 publications, il est préférable d'avoir 1000 publications qui sont fortement en lien avec la requête effectuée, c'est-à-dire avec la maladie rare dont on cherche les publications. Ce paramètre se nomme « **sort** » et il suffit d'ajouter **sort=relevance** dans nos requêtes. Les autres paramètres sont laissés à leur valeur par défaut.

Dans ce travail, on supposera que les publications prélevées grâce à ce paramétrage seront véritablement en lien avec la maladie recherchée. En ce qui concerne la valeur « **relevance »** du paramètre « **sort** », nous avons de bonnes raisons de penser que le tri réalisé par l'API est réellement pertinent au vu de ce qui est décrit dans les liens suivants :

https://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp. Algorithm_for_finding_best_ma

https://dataguide.nlm.nih.gov/eutilities/utilities.html

Pour résumé, l'établissement des scores de pertinence de chaque publication est d'abord réalisé par une variante de TF-IDF. Ensuite, une fois les publications classées avec ces scores, les publications ayant le score le plus élevé sont retriées en fonction d'un nouveau score établi cette fois-ci par un algorithme de « machine learning » qu'on appelle « forest of gradient boosted trees ». Cet algorithme est en réalité une méthode d'agrégation de modèles utilisant de le gradient de la fonction de perte pour la construction de chaque modèle (Burges, 2010). L'utilisation de cet algorithme permet une amélioration significative des performances sur les résultats de recherche.

4.3.4.4 STOCKAGE DES PUBLICATIONS

Le stockage des publications a d'abord été réalisé avec une base de données relationnelle MySQL puis nous avons changé de type de stockage en passant à MongoDB, un système de base de données orienté documents, plus adapté à notre problème de stockage. La totalité des publications est stockée dans une collection nommée « Publication », chaque publication possédant un champ faisant une référence à la maladie qui lui est liée.

4.3.4.5 PROBLÈMES ET DIFFICULTÉS RENCONTRÉS

En dehors de la problématique de stockage, nous avons eu un certain questionnement concernant l'exploitation des publications. La question était de savoir si nous avions réellement besoin de stocker les publications, car une alternative reste possible. Elle consiste en le fait de réaliser les requêtes sur PMC et d'immédiatement traiter et faire notre extraction sur les résultats renvoyés par la requête sans passer par une phase de stockage. Cette approche a l'avantage d'être plus rapide, car il n'y a pas de stockage à effectuer, mais elle implique qu'à chaque fois que l'on a besoin du texte des publications, il est nécessaire de faire une requête. Or nous voulions avoir un accès rapide au texte pour lui appliquer de potentielles transformations (par exemple de la tokenisation, du « stemming »…) ; sans oublier que nous nous sommes limités à 1000 publications par maladie (ce qui représente une quantité de données suffisamment petite pour stocker facilement). Tout cela fait en sorte que le stockage par nos propres moyens dans une base de données est le meilleur choix.

4.3.5 « HUMAN PHENOTYPE ONTOLOGY »

La « Human Phenotype Ontology » (HPO) (Köhler et al., 2017) est une ontologie proposant un vocabulaire standardisé des anomalies phénotypiques chez l'humain. Une ontologie est une représentation informatique d'un domaine de connaissance décrivant les différentes entités appartenant à ce domaine ainsi que leurs relations entre elles. Le terme relation peut être interprété de différentes manières, mais dans le cas de HPO, le seul type de relation utilisé est « *is_a* » (Relations simples de classe-sous-classe). Par exemple, « *Abnormality of the feet* » *is_a* « *Abnormality of the lower limbs* »

HPO possède 5 sous-ontologies, à savoir :

- « Phenotypic Abnormality » : Ontologie principale regroupant l'ensemble des anomalies phénotypiques
- « Mode of Inheritance » : Regroupe les modes d'héritage des anomalies phénotypiques (génétique, sporadique, multifactoriel…)
- « Clinical modifier » : Regroupe les termes permettant de caractériser les anomalies phénotypiques en fonction de la sévérité, l'âge d'apparition, latéralité et d'autres aspects
- « Clinical course » : Cette sous-ontologie décrit l'évolution typique d'une maladie dès son apparition, sa progression dans le temps et la résolution éventuelle ou la mort de l'individu atteint.
- « Frequency » : Regroupe les termes représentatifs de la fréquence des anomalies phénotypiques (fréquent, rare, très rare, très fréquent…)

Dans notre cas, nous utiliserons seulement l'ontologie principale « Phenotypic Abnormality », car nous utilisons HPO comme une simple liste d'anomalies phénotypiques, nous ne considérons pas les autres aspects.

Il faut savoir que cette terminologie est également utilisée pour détecter des anomalies phénotypiques dans les publications. L'utilisation d'une telle terminologie à la fois dans la phase de détection dans les publications et dans le jeu de vérification permet de s'assurer d'effectuer une évaluation cohérente des performances.

4.4 ALGORITHMES UTILISÉS

Une fois les sources de données trouvées, nous avons dû mettre en place les algorithmes qui vont exploiter ces données et ainsi réaliser l'extraction de nos éléments du texte des publications scientifiques.

4.4.1 RECONNAISSANCE BASÉE SUR UN DICTIONNAIRE

Afin d'extraire nos anomalies phénotypiques, nous avons choisi d'utiliser l'approche par dictionnaire vu en 3.2.3.1.2. Elle consiste en la recherche de mots d'un dictionnaire (le dictionnaire étant HPO) dans les publications scientifiques : quand on trouve une correspondance, on extrait le terme comme une anomalie phénotypique appartenant à la maladie liée à la publication.

Nos premiers résultats, ayant donné, comme prévu, un grand nombre de faux positifs. Nous nous sommes questionnés sur le fait d'utiliser l'approche « Approximate Dictionnary-based », mais il se trouve qu'elle avait pour effet d'augmenter plus le nombre de faux positifs que de vrais positifs. Son utilisation ne semblait alors pas être une bonne idée. Au final, l'approche par dictionnaire a été réalisée par l'utilisation de la librairie LingPipe par Alias-i (2008) qui propose une implémentation de cette approche.

4.4.2 TF-IDF MODIFIÉ

Comme dit précédemment, l'approche par dictionnaire donne un grand nombre de faux positifs (c'est-à-dire que trop d'éléments non pertinents ont été extraits). Au vu de cette observation, nous avons voulu tenter une approche pour éliminer le surplus d'éléments pour garder un maximum de vrais positifs. Notre idée est la suivante : pour chaque maladie, attribuons une valeur d'importance à chaque entité extraite, classons-les par valeur décroissante et plaçons un seuil d'importance général pour éliminer les entités qui ne sont pas assez « importantes ». De cette idée découlent deux choses à déterminer :

- Le calcul de la valeur d'importance
- Un seuil optimal pour éliminer le maximum d'entités « non importantes » tout en conservant le plus d'entités « importantes »

Dans cette section, nous parlerons du calcul de la valeur d'importance qui est basé sur la statistique TF-IDF. L'exemple qui suit permettra d'illustrer le fonctionnement de notre TF-IDF modifié : soit les maladies suivantes : « Extraneural perineurioma » (identifiable sur Orphanet par l'identifiant suivant : ORPHANUMBER : 100002) et « Ocular albinism with late-onset sensorineural deafness » (ORPHANUMBER : 1000). Et soit les anomalies phénotypiques suivantes : « neoplasm », « pain », « hearing impairment », « breast carcinoma ».

Ces anomalies ont été trouvées dans les publications scientifiques de PMC en rapport avec ces maladies et plus précisément :

- « neoplasm » a été trouvé dans des publications des deux maladies
- « pain » a été trouvé dans les publications de la maladie 100002
- « hearing impairment » a été trouvé dans les publications de la maladie 1000
- « breast carcinoma » a été trouvé dans les publications de la maladie 1000

Aussi, avant de continuer, nous allons poser la notation qui sera utilisée pour bien comprendre nos expressions mathématiques. Notre but est d'identifier les termes ***t*** parmi l'ensemble des termes correspondant à des anomalies phénotypiques ***T*** dans des publications ***p*** en rapport avec des maladies ***m***. De cette manière, ***M*** désigne l'ensemble des maladies rares, ***P*** désigne l'ensemble des corpus de publications de toutes les maladies. Sa formule est la suivante :

$$\boldsymbol{P} = \{\{\boldsymbol{p}\}_{\boldsymbol{m}} | \boldsymbol{m} \in \boldsymbol{M}\} \quad (4)$$

où $\{p\}_m$ désigne le corpus de publications p liées à la maladie m

4.4.2.1 TF (« TERM FREQUENCY »)

On reprend le principe du premier facteur de TF-IDF : TF, désignant « *Term Frequency* », représentait la fréquence d'un mot dans un document. Dans notre cas, le principe est similaire, mais appliqué sur toutes les publications d'une maladie ; pour simplifier, toutes les publications en rapport avec une maladie deviennent un seul et même document. On a donc ici, notre nouvelle « Term Frequency » (***TF(t, m)***), qui désigne la fréquence d'un terme dans l'ensemble des publications d'une maladie. Par exemple, en prenant la fréquence brute, cela correspond au nombre de fois où le mot « neoplasm » apparaît dans toutes les publications de la maladie « Extraneural perineurioma ». Cette fréquence brute, notée $\boldsymbol{f_{t,m}}$, représente le nombre d'occurrences du terme ***t*** dans l'ensemble des

publications $\{\boldsymbol{p}\}_{\boldsymbol{m}}$ de la maladie $\boldsymbol{m}$. Les différentes versions que nous utiliserons de cette nouvelle TF seront les suivantes :

TABLEAU 8: FORMULES DES DIFFÉRENTES VERSIONS DE NOTRE « TERM FREQUENCY »

Version	Nom utilisé dans les graphiques	Formulation
Binaire	Binary	{0, 1}
Compte brut	RawCount	$\boldsymbol{f_{t,m}}$
Normalisation logarithmique	LogNorm	$\boldsymbol{log}(\mathbf{1} + \boldsymbol{f_{t,m}})$
Normalisation double 0	MinMaxNorm	$\frac{\boldsymbol{f_{t,m}}}{\boldsymbol{max}_{\{t' \in P\}} \boldsymbol{f_{t',m}}}$

où $\boldsymbol{max}_{\{t' \in P\}} \boldsymbol{f_{t',m}}$ désigne le $\boldsymbol{f_{t,m}}$ maximal parmi l'ensemble des $\boldsymbol{f_{t,m}}$ de la maladie **m**

4.4.2.2 IDF (« INVERSE DISEASE FREQUENCY »)

De même, pour le second facteur, IDF, on va reprendre son principe général qui est, pour rappel, un facteur compensant TF par la fréquence dans l'ensemble des documents. Dans notre cas, au lieu de considérer un ensemble de documents, on va plutôt considérer un ensemble de maladies. Ainsi, l'« Inverse Document Frequency » devient l'« Inverse Disease Frequency ». L'idée est de créer un facteur qui a le comportement suivant : plus un terme est présent dans l'ensemble des publications de maladies, plus le facteur sera petit, amoindrissant.

Pour ce faire, il est important de bien définir le fait qu'un terme soit **présent**, c'est-à-dire définir sa fréquence intermaladie. Nous avons retenu deux définitions différentes pour définir sa fréquence intermaladie (sa fréquence dans l'ensemble de maladies). Ces deux fréquences seront testées et discutées en 4.6. Premièrement, la fréquence intermaladie 1 ($\boldsymbol{f_{inter1,t}}$) désigne le nombre de maladies pour lesquelles le terme **t** apparaît dans l'ensemble des publications $\{\boldsymbol{p}\}_{\boldsymbol{m}}$, soit mathématiquement :

$$\boldsymbol{f_{inter1,t}} = |\{\boldsymbol{m} \in \boldsymbol{M} : \boldsymbol{t} \in \{\boldsymbol{p}\}_{\boldsymbol{m}}\}| \quad (5)$$

Deuxièmement, la fréquence intermaladie 2 ($\boldsymbol{f_{inter2,t}}$) désigne la somme, parmi l'ensemble des maladies $\boldsymbol{M}$, des TF liées au terme **t**, soit mathématiquement :

$$\boldsymbol{f_{inter2,t}} = \sum_{\boldsymbol{m} \in \boldsymbol{M}} \boldsymbol{TF}(\boldsymbol{t}, \boldsymbol{m}) \quad (6)$$

Cette deuxième fréquence tente d'ajouter une notion supplémentaire à la première. Là où la première ajoute 1.0 à chaque fois qu'un terme se trouve dans les publications d'une maladie, la deuxième tente d'ajouter une représentation pondérée de la présence du terme dans les publications de la maladie, soit TF.

On remarquera que si on fait le choix d'utiliser la TF de type « Binaire » (vu en 3.2.4.3) dans la formule de la deuxième fréquence, on obtient une mesure identique à la première fréquence. Les différentes versions d'IDF qui seront utilisées seront les suivantes :

TABLEAU 9: FORMULES DES DIFFÉRENTES VERSIONS DE NOTRE « INVERSE DOCUMENT FREQUENCY »

Version	Nom utilisé dans les graphiques	Formulation
Unaire	Unary	1
Fréquence intermaladie 1	NbDisease_i	$\boldsymbol{f_{inter1,t}}$
Inverse fréquence intermaladies 1	Inverse_NbDisease_i	$\frac{\lvert\boldsymbol{M}\rvert}{\boldsymbol{f_{inter1,t}}}$
IDF_1	IDF_Classic_NbDisease_i	$\boldsymbol{log}\left(\frac{\lvert\boldsymbol{M}\rvert}{\boldsymbol{f_{inter1,t}}}\right)$
IDF_1 Smooth	IDF_Smooth_NbDisease_i	$\boldsymbol{log}\left(\mathbf{1}+\frac{\lvert\boldsymbol{M}\rvert}{\boldsymbol{f_{inter1,t}}}\right)$
Probabilistic IDF_1	Prob_IDF_NbDisease_i	$\boldsymbol{log}\left(\frac{\lvert\boldsymbol{M}\rvert-\boldsymbol{f_{inter1,t}}}{\boldsymbol{f_{inter1,t}}}\right)$
Fréquence intermaladie 2	SumOfMinMaxNorm_i	$\boldsymbol{f_{inter2,t}}$
Inverse fréquence intermaladies 2	Inverse_SumOfMinMaxNorm_i	$\frac{\boldsymbol{SomTot}}{\boldsymbol{f_{inter2,t}}}$
IDF_2	IDF_Classic_SumOfMinMaxNorm_i	$\boldsymbol{log}\left(\frac{\boldsymbol{SomTot}}{\boldsymbol{f_{inter2,t}}}\right)$
IDF_2 Smooth	IDF_Smooth_SumOf-MinMaxNorm_i	$\boldsymbol{log}\left(\mathbf{1}+\frac{\boldsymbol{SomTot}}{\boldsymbol{f_{inter2,t}}}\right)$
Probabilistic IDF_2	Prob_IDF_SumOfMinMaxNorm_i	$\boldsymbol{log}\left(\frac{\boldsymbol{SomTot}-\boldsymbol{f_{inter2,t}}}{\boldsymbol{f_{inter2,t}}}\right)$

où $\boldsymbol{SomTot}$ est la somme, parmi l'ensemble des termes T correspondant à des anomalies phénotypiques, des $\boldsymbol{f_{inter2,t}}$ dont la formule est la suivante :

$$\boldsymbol{SomTot} = \sum_{t \in T} \boldsymbol{f_{inter2,t}} \quad (7)$$

Les formules ont été créées et choisies à partir des différentes variantes existantes pour un facteur IDF classique (vu en 3.2.3.2.2).

4.4.2.3 UTILISATION DE LA NOUVELLE TF-IDF

La nouvelle mesure de score d'importance servira à classer les éléments extraits de manière décroissante. Étant donné les différentes versions qui ont été présentées, nous allons tester chaque combinaison de couple (TF, IDF) afin de trouver la plus performante. La plus performante sera définie par celle qui tendra à attribuer des valeurs hautes aux éléments extraits qui sont réellement importants. (Voir 4.6.1.3 pour cette performance). Une fois le choix de la combinaison effectué, il restera à déterminer le seuil optimal ; la méthodologie pour le trouver sera détaillée en 4.5.4 et sa valeur sera donnée en 4.6.4.4.

4.5 OUTIL DÉVELOPPÉ

Les sources de données et les algorithmes choisis, un système a été implémenté pour avoir une réalisation concrète. Ainsi, la section qui suit expliquera le fonctionnement global du système en proposant des explications et des schémas de chaque étape.

4.5.1 RÉCUPÉRATION DES PUBLICATIONS ET NER PAR DICTIONNAIRE

La Figure 9 ci-dessous résume le processus de récupération des publications ainsi que leur exploitation par le système de NER par dictionnaire. Numérotée de 1 à 6, elle présente les différentes étapes du processus jusqu'à l'utilisation de HPO pour faire notre NER.

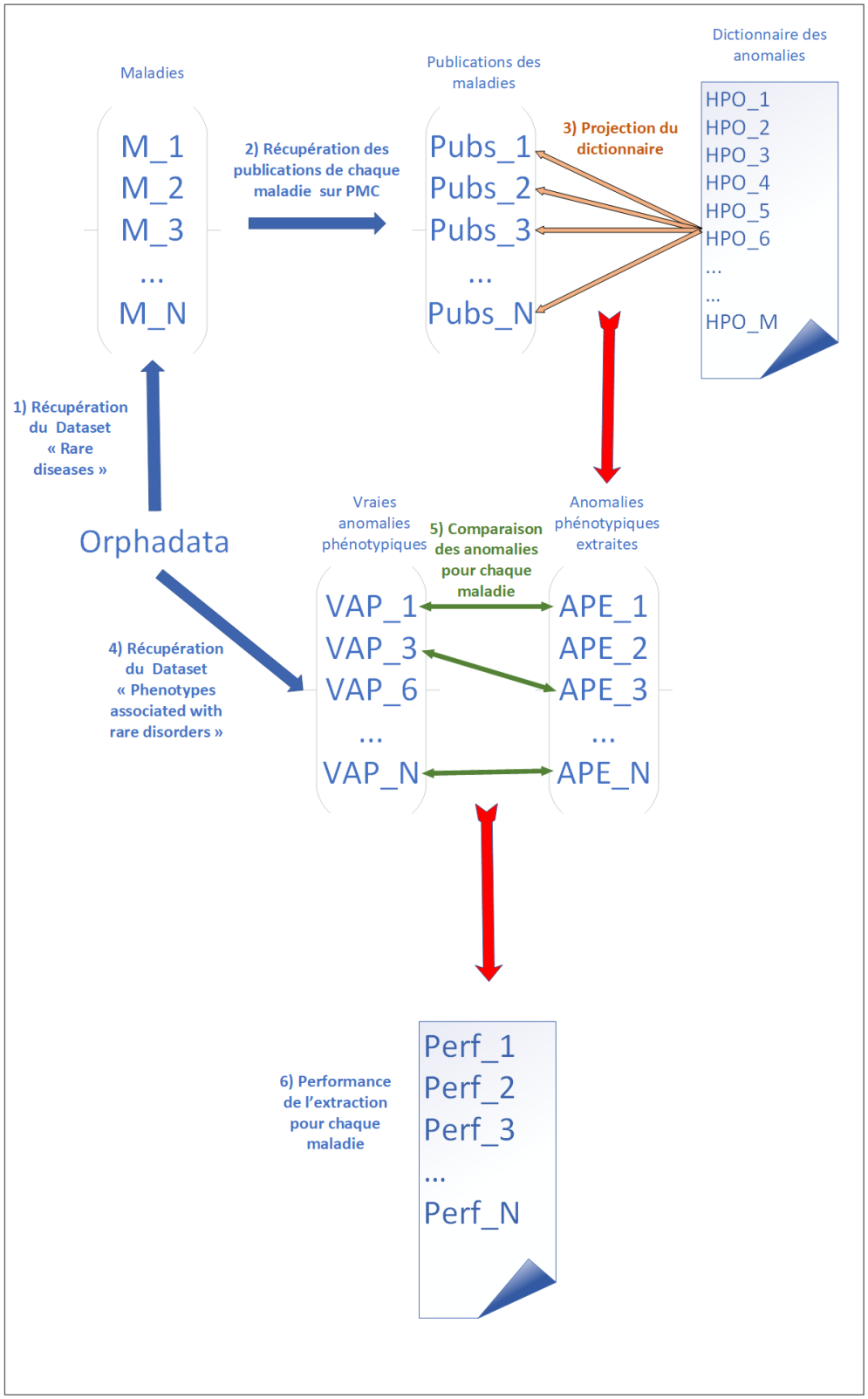

FIGURE 9: SCHÉMA DES ÉTAPES DE RÉCUPÉRATION DES PUBLICATIONS ET NER PAR DICTIONNAIRE

La première étape est de récupérer les maladies rares ainsi que leurs informations de base à partir d'OrphaData (Étape 1). Ensuite, on s'occupe de récupérer les publications liées à cette maladie rare en allant les chercher sur PMC (Étape 2). Ces publications subissent ensuite une étape de projection du dictionnaire HPO, c'est-à-dire qu'on cherche des occurrences des termes du dictionnaire dans les publications et on les sauvegarde s'ils sont repérés (Étape 3). Après avoir réalisé notre extraction, on prépare le jeu de données qui va nous servir pour faire une première évaluation des éléments extraits, c'est-à-dire qu'on cherche les « Véritables anomalies phénotypiques » pour les confronter à nos « Anomalies phénotypiques extraites » (Étape 4). Enfin, après avoir faire la comparaison entre les deux (Étape 5), on récupère les mesures de performances calculées (Étape 6) pour chaque maladie.

4.5.2 PREMIÈRE ÉVALUATION

Quand l'étape 4 est terminée, le but est de pouvoir comparer les informations extraites avec les « véritables » informations. Plus ces deux jeux de données seront proches, plus notre extraction sera de bonne qualité. La Figure 10 détaille le fonctionnement pour la comparaison concernant une maladie i avec **VAP_i** désignant les véritables anomalies phénotypiques de la maladie i et **APE_i** les anomalies phénotypiques extraites de la maladie i

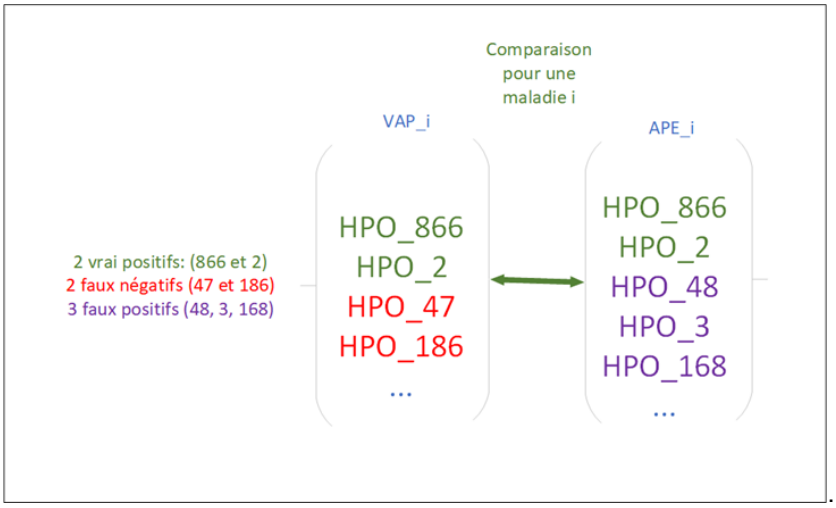

FIGURE 10: SCHÉMA DE LA PREMIÈRE ÉVALUATION

Le détail de l'évaluation et des mesures de performances sera expliqué dans la section sur les expérimentations (4.6).

4.5.3 RECHERCHE DE LA MEILLEURE COMBINAISON TF-IDF

Comme présenté en 4.4.2, nous allons utiliser notre version modifiée de TF-IDF pour calculer une valeur d'importance de chaque anomalie phénotypique extraite. Cependant, la TF et l'IDF possédant chacune différentes versions, il est nécessaire de faire un choix parmi ces versions. Nous avons choisi de faire le choix des versions en fonction d'une mesure de performance caractérisant la segmentation entre les véritables anomalies phénotypiques (vrais positifs) et les anomalies phénotypiques extraites « en trop » (faux positifs). Cette mesure, que nous allons appeler *rang moyen des « véritables » anomalies phénotypiques* a pour objectif de donner le rang moyen des véritables anomalies phénotypiques parmi l'ensemble des anomalies phénotypiques extraites classées dans l'ordre décroissant de valeur d'importance pour une maladie donnée.

Ainsi, si le rang moyen est proche de 1, on pourra en déduire que les « véritables » anomalies phénotypiques ont tendance à avoir une valeur d'importance haute et donc que la combinaison de TF-IDF choisie sépare bien les « bonnes » des « mauvaises » anomalies phénotypiques (en plaçant les « bonnes » en début de liste et les « mauvaises » en fin de liste).

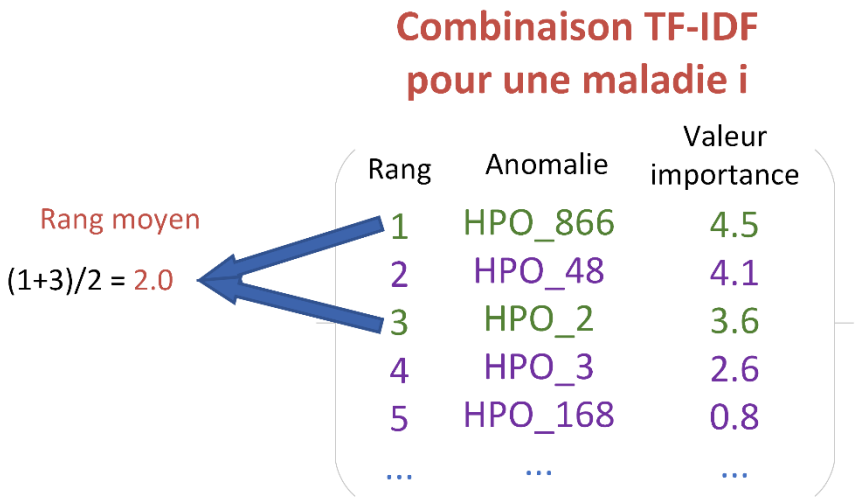

FIGURE 11: ILLUSTRATION DU CALCUL DU RANG MOYEN DES VÉRITABLES ANOMALIES PHÉNOTYPIQUES POUR UNE MALADIE I (LES ANOMALIES SONT CLASSÉES PAR VALEUR D'IMPORTANCE DÉCROISSANTE)

Une fois les rangs moyens calculés pour toutes les maladies pour une combinaison de TF-IDF, on en fait la moyenne (qu'on appellera RANG MOYEN) et on l'attribue à la combinaison. On

répète alors l'opération pour toutes les combinaisons possibles de TF-IDF (11 IDF x 4 TF = 44 combinaisons) pour obtenir tous les RANGS MOYENS. La combinaison possédant le RANG MOYEN le plus petit est considérée comme la meilleure combinaison. La Figure 11 illustre le calcul du RANG MOYEN d'une maladie i avec en vert, les véritables anomalies phénotypiques (vrais positifs) et en violet, les anomalies phénotypiques extraites en surplus (faux positifs).

4.5.4 RECHERCHE DU MEILLEUR SEUIL

Une fois la meilleure combinaison TF-IDF déterminée, comme mentionnée à la section 4.4.2, la prochaine préoccupation est de parvenir à déterminer un seuil optimal. Pour être plus concret, toutes les anomalies phénotypiques ayant une valeur d'importance inférieure au seuil sont éliminées. Un seuil sera considéré comme optimal s'il parvient, en moyenne sur toutes les maladies, à éliminer les anomalies phénotypiques en surplus tout en conservant les véritables anomalies phénotypiques.

Par exemple, si on reprend les anomalies extraites dans la Figure 11, en prenant un seuil de 3.5, on se retrouverait avec 2 véritables anomalies et 1 seule anomalie en surplus. Au contraire, si on prend un seuil trop bas comme 0,7, on se retrouverait avec toujours 2 véritables anomalies, mais également avec 3 anomalies en surplus ; ce qui est clairement moins bon qu'en utilisant le premier seuil de 3,5.

L'approche envisagée pour la recherche de ce seuil optimal est une approche simple qui consiste à tester un ensemble de valeurs possibles dans un intervalle donné. Par exemple, en prenant l'intervalle [0.0, 5.0] et un pas de 0.01, l'ensemble des valeurs de seuil testées sera de la forme {0.0, 0.01, 0.02, 0.03, …, 4.98, 4.99, 5.00}. Nous avons opté pour cette approche simple, car lors de nos essais, nous avons vu qu'elle pouvait se réaliser dans un temps raisonnable. En ce qui concerne le choix de l'intervalle et du pas, nous avons procédé par essais successifs, en tentant à chaque fois de cibler la zone présentant les meilleures performances, c'est-à-dire en augmentant le pas et en diminuant l'intervalle dans vers cette zone.

4.5.5 IMPLÉMENTATION

L'implémentation de notre système a été réalisée en C# avec l'environnement de développement Visual Studio 2017. Le système est sous la forme d'une solution Visual Studio contenant

différents sous projets, chacun ayant une responsabilité bien distincte. Le stockage des données des maladies, des publications, des éléments extraits est fait dans une base de données MongoDB avec, respectivement, les collections « Disease », « Publication », « PredictionDataRepository ». En ce qui concerne les données de comparaison, c'est-à-dire celles qui permettent de procéder à l'évaluation des performances, la collection se nomme « RealDataRepository ». Le code a été versionné à l'aide de GIT et est disponible sur le système de gestion de développement en ligne Github à l'adresse https://github.com/CharlesCousyn/RDSearch4 .

4.6 EXPÉRIMENTATIONS

Après avoir parlé du fonctionnement du système d'extraction et de sa réalisation, il est temps de parler de la qualité des éléments extraits et de la manière de l'évaluer. Nous commencerons par parler des mesures de performances pertinentes.

4.6.1 MESURES DE PERFORMANCES

Ce qui est important dans le choix d'une ou plusieurs mesures de performance, c'est d'abord de savoir préciser l'objectif de notre système. Dans notre cas, dire que le système est performant revient à dire qu'il extrait un grand nombre de véritables anomalies phénotypiques, peu d'anomalies en surplus et peu d'anomalies oubliées.

4.6.1.1 PRÉCISION ET RAPPEL

La précision et le rappel sont deux mesures largement utilisées et qui sont capables de répondre à notre problème de performance. Avant de parler de celles-ci, il est nécessaire de poser les définitions suivantes pour une maladie $\boldsymbol{m}$ de l'ensemble des maladies $\boldsymbol{M}$:

- **Vrai positif $\boldsymbol{m}$** : Anomalie phénotypique prédite et se trouvant dans les « véritables » anomalies phénotypiques de la maladie $\boldsymbol{m}$
- **Faux positif $\boldsymbol{m}$** : Anomalie phénotypique prédite et ne se trouvant pas dans les « véritables » anomalies phénotypiques de la maladie $\boldsymbol{m}$ (anomalies en surplus)
- **Faux négatif $\boldsymbol{m}$** : Anomalie phénotypique non prédite et se trouvant dans les « véritables » anomalies phénotypiques de la maladie $\boldsymbol{m}$ (anomalies oubliées)

- **Vrai négatif $\boldsymbol{m}$** : Anomalie phénotypique non prédite et ne se trouvant pas dans les « véritables » anomalies phénotypiques de la maladie $\boldsymbol{m}$

À partir de ces définitions, on peut alors donner les formules de la précision et du rappel pour une maladie $\boldsymbol{m}$ donnée :

$$Précision_m = \frac{Vrai\ positifs\ m}{Vrai\ positifs\ m + Faux\ positifs\ m} \quad (8)$$

$$Rappel_m = \frac{Vrai\ positifs\ m}{Vrai\ positifs\ m + Faux\ Négatifs\ m} \quad (9)$$

La quantité « **Vrai positifs $\boldsymbol{m}$ + Faux positifs $\boldsymbol{m}$** » représentent le nombre d'anomalies phénotypiques prédites par notre approche alors que la quantité « **Vrai positifs $\boldsymbol{m}$ + Faux négatifs $\boldsymbol{m}$** » représente le nombre de véritables anomalies phénotypiques pour la maladie $\boldsymbol{m}$.

En ce qui concerne les vrais négatifs, on remarquera que nous ne l'avons pas utilisé. La raison est simple : la notion de vrai négatif n'a pas vraiment de sens, pour la simple raison que, dans notre cas, notre jeu de vérification ne contient que les éléments « positifs » et que quelque chose qui est considéré comme « négatif » est en réalité absent de ce jeu de données.

On notera que ces mesures concernant une seule maladie, d'autres mesures plus globales sont utilisées pour évaluer l'ensemble des maladies. Ces mesures ont pour formules :

$$Précision\ Moyenne = \frac{\sum_{m \in M} Vrai\ positifs_m}{\sum_{m \in M} Vrai\ positifs_m + \sum_{m \in M} Faux\ positifs_m} \quad (10)$$

$$Rappel\ Moyen = \frac{\sum_{m \in M} Vrai\ positifs_m}{\sum_{m \in M} Vrai\ positifs_m + \sum_{m \in M} Faux\ négatifs_m} \quad (11)$$

4.6.1.2 F-MESURE OU F-SCORE

La F-mesure ou F-score est une mesure qui prend en considération la précision et le rappel, c'est en réalité leur moyenne harmonique, elle sera utilisée pour savoir si globalement, notre précision et notre rappel sont bons ou non pour une maladie $\boldsymbol{m}$. Voici sa formule :

$$F - Score_m = 2 \times \frac{Précision_m \times Rappel_m}{Précision_m + Rappel_m} \quad (12)$$

Avec également sa version plus globale pour évaluer l'ensemble des maladies :

$$F - Score\ Moyen = 2 \times \frac{Précision\ Moyenne \times Rappel\ Moyen}{Précision\ Moyenne + Rappel\ Moyen} \quad (13)$$

4.6.1.3 RANG MOYEN DES « VÉRITABLES » ANOMALIES PHÉNOTYPIQUES

Comme dit en 4.5.3, une autre mesure a été utilisée au vu de nos premiers résultats. Dans notre approche, chaque anomalie phénotypique extraite possède une valeur d'importance, un poids. Une fois cette valeur calculée, nous l'utilisons pour trier les anomalies phénotypiques extraites dans l'ordre décroissant. (Plus la valeur est grande, plus l'élément est important et plus son rang se rapproche de 1). Ce poids permet donc d'obtenir un rang dans le classement. En posant :

n_m : $le\ nombre\ de\ vraies\ anomalies\ phénotypiques\ pour\ la\ maladie\ m$,

RVP_{im} : $le\ rang\ de\ la\ vraie\ anomalie\ phénotypique\ i\ pour\ la\ maladie\ m$,

$RMVP_m$: $le\ rang\ moyen\ des\ vraies\ anomalies\ phénotypiques\ pour\ la\ maladie\ m$,

M : $l'ensemble\ des\ maladies\ rares$

$$\boldsymbol{RMVP_m} = \frac{\sum_{i=1}^{\boldsymbol{n_m}} \boldsymbol{RVP_{im}}}{\boldsymbol{n_m}} \quad (14)$$

Cette mesure concernant une seule maladie, nous avons une mesure qui réalise la moyenne pour toutes les maladies :

$$\boldsymbol{RMVP} = \frac{\sum_{m \in M} \boldsymbol{RMVP_m}}{|\boldsymbol{M}|} \quad (15)$$

4.6.2 FORMAT DES RÉSULTATS

Les résultats de l'évaluation sont tous générés au format JSON, car ce format permet de représenter facilement la totalité des mesures de performances.

Il existe trois formats de résultats. Le premier format, dont les noms des fichiers générés sont de la forme « ***results_[timestamp].json*** » donne les performances générales ainsi que les performances par maladie (***[timestamp]*** représente l'horodatage, soit l'instant auquel le fichier de résultats a été généré). Le second format, dont les noms de fichiers sont de la forme « ***metaResults_[timestamp].json*** », donne seulement les performances générales (pour toutes les maladies), mais cette

fois, en utilisant différentes formules pour calculer la valeur d'importance des anomalies phénotypiques. On se retrouve donc avec un ensemble de performances générales. Enfin, le dernier format, lui, a des fichiers nommés sous la forme « ***metaResultsWeigth_[timestamp]***.*json* ». Ce format donne lui aussi les performances générales, mais cette fois, en utilisant des seuils différents pour éliminer des anomalies en surplus. À chaque évaluation, un nouveau fichier est généré avec un nom différent grâce à la présence du timestamp. On peut ainsi garder un historique des résultats.

4.6.3 OUTIL DE VISUALISATION

Une fois ces fichiers de résultats générés, il est alors intéressant d'avoir un outil capable de visualiser le contenu de ces fichiers à travers l'utilisation de graphiques. En effet, les fichiers peuvent faire jusqu'à 33000 lignes, ce qui devient difficilement exploitable pour un humain ; il est alors plus simple pour un programme de lire ces données et de les mettre sous une forme plus simplifiée, qu'un humain pourra interpréter. L'outil en question a été développé, des captures d'écran seront montrées dans la prochaine section. Pour plus de détails, l'outil est disponible à l'adresse suivante https://github.com/CharlesCousyn/GraphJSON-for-RDSearch .

4.6.4 RÉSULTATS ET INTERPRÉTATION

4.6.4.1 PERFORMANCES DE L'APPROCHE PAR DICTIONNAIRE SEULE

Les performances qui vont suivre sont les résultats obtenus quand l'approche par dictionnaires est utilisée sans l'utilisation de seuil. On s'attend donc à une présence d'un grand nombre de faux positifs, soit à une précision très basse. Le Tableau 10 regroupe les résultats moyens sur l'ensemble des maladies.

TABLEAU 10:TABLEAU DE RÉSULTATS MOYENS DE L'APPROCHE PAR DICTIONNAIRE SEULE

Précision moyenne	Rappel moyen	F-score moyen
1.83%	59.66%	3.55%

On notera que le rappel est, en moyenne, assez élevé, cela est dû à l'approche par dictionnaire qui vérifie pour chaque mot de chaque publication, si c'est une anomalie phénotypique ou non.

Si le rappel ne dépasse pas 60% sur toutes les maladies, cela signifie qu'il manque, en moyenne, presque 40% des anomalies phénotypiques dans les publications. Notre hypothèse est qu'il doit manquer des publications à analyser dans notre base de données et que PMC ne contient pas tout le contenu nécessaire sur les maladies rares. Plus concrètement, cela signifie aussi que notre extraction ne pourra jamais avoir un rappel supérieur à 60% et donc que, mathématiquement, le F-score est limité à 75%.

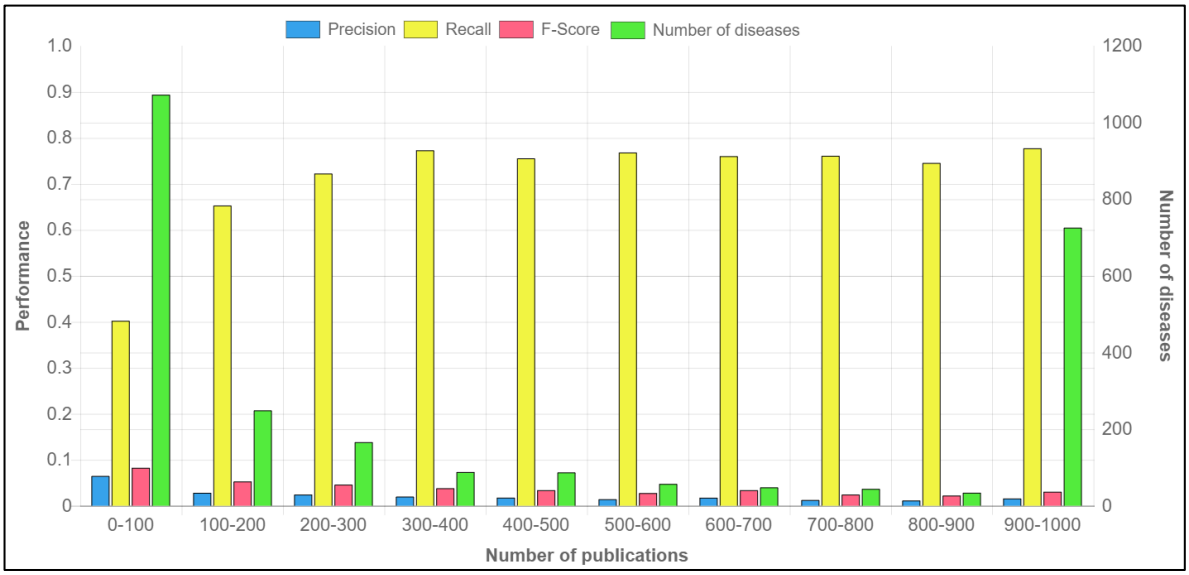

FIGURE 12: GRAPHIQUE DE VISUALISATION DES FICHIERS "RESULTS"

La Figure 12 est la visualisation du premier type de fichier (« *results_[**timestamp**].json* »), il représente les différentes mesures de performances de notre extraction (toujours sans utilisation de seuil) en fonction du nombre de publications. Nous avons fait ce choix, car nous voulions tester l'impact de la quantité de publications sur les performances. On peut noter plusieurs choses.

Au total, 2572 maladies ont été finalement évaluées (sommes des barres vertes). C'est tout à fait attendu, car notre jeu de données de prédiction et notre jeu de données réelles sur les anomalies phénotypiques ne possèdent pas d'informations sur toutes les maladies. Pour être exact, il y a 3056 maladies dont on connaît les véritables anomalies (jeu de données réelles) et il y a 5894 maladies qui possèdent au moins une publication dans notre base de données (jeu de données de prédiction) et en croisant ces deux ensembles, on se retrouve avec 2572 maladies en commun entre les jeux de données.

Ensuite, il y a quelques remarques à faire sur le nombre de publications des maladies rares. De nombreuses maladies ont entre 0 et 100 publications (1073 pour être exact). Entre 0 et 900 publications, on remarque que plus le nombre de publications augmente, moins il y a de maladies concernées. Cela confirme qu'au moins une portion des maladies rares est peu documentée. Enfin 726 maladies possèdent au moins 900 publications (1000 est un plafond fixé par nous-mêmes pour des questions de stockage).

En ce qui concerne le rappel, on voit d'abord qu'il est assez élevé en général, allant jusqu'à 78.44% et qu'il a tendance à monter avec le nombre de publications **(1)**. En ce qui concerne la précision, on voit tout d'abord qu'elle est très faible, entre 1% et 7%, comme attendu, qu'elle a tendance à baisser avec le nombre de publications **(2)**. Enfin le F-score, qui est le reflet de ces deux mesures combinées, tend à baisser avec le nombre publications et reste très faible, entre 2% et 8%.

Les observations **(1)** et **(2)** sont parfaitement normales aux vues de l'approche utilisée et de la définition de notre rappel et de notre précision. En effet, l'approche par dictionnaire permet d'obtenir un nombre faible de faux négatifs (Anomalies phénotypiques non prédites et se trouvant dans les « véritables » anomalies phénotypiques), mais un nombre élevé de faux positifs (Anomalies phénotypiques prédites et ne se trouvant pas dans les « véritables » anomalies phénotypiques).

En résumé, cette approche permet de trouver presque toutes les véritables anomalies phénotypiques, mais en contrepartie, elle trouve aussi beaucoup d'anomalies phénotypiques qui ne sont pas liées à la maladie.

4.6.4.2 CHOIX DE LA COMBINAISON TF-IDF

La Figure 13 est une visualisation du second type de fichier (« *metaResults_[**timestamp**].json* »), elle montre le rang moyen des vraies anomalies phénotypiques en fonction de la « Term Frequency » et de l'IDF utilisées (classement fait en fonction de TF*IDF décroissant pour rappel). Voici quelques indications pour bien comprendre ce graphique : un point petit correspond à un bon rang moyen (Rang moyen faible), un point gros correspond à un mauvais rang moyen (Rang moyen élevé), un point rouge correspond à un mauvais écart type sur la mesure du rang moyen

(Écart type élevé) et un point bleu correspond à un bon écart type sur la mesure du rang moyen (Écart type faible).

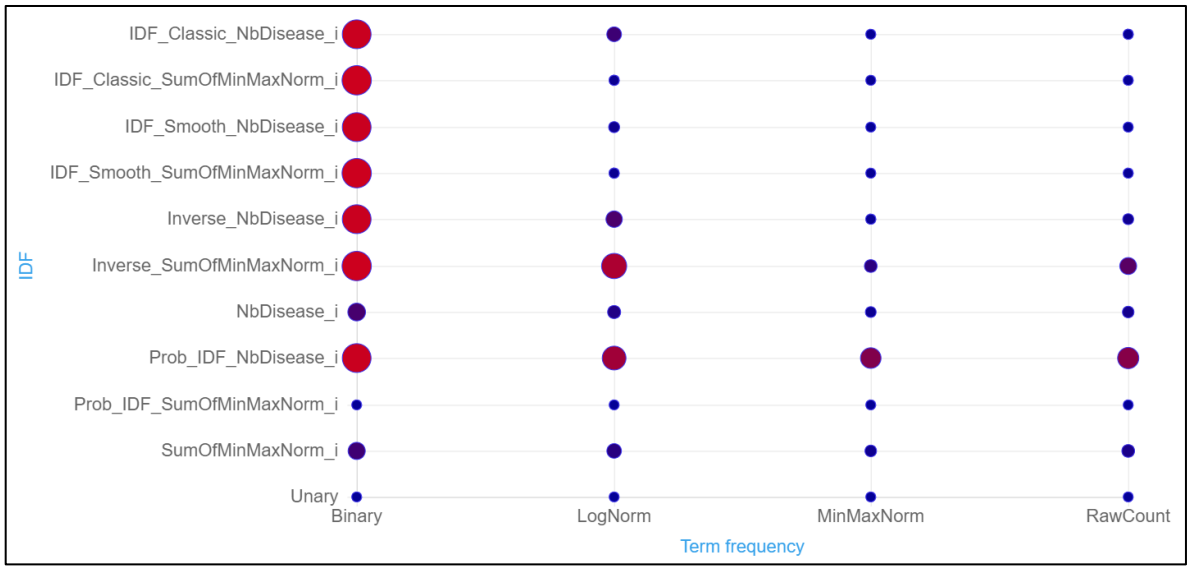

FIGURE 13: GRAPHIQUE DE VISUALISATION DES FICHIERS "METARESULTS"

Plusieurs choses découlent de l'observation du graphe. D'abord, on remarquera que globalement, la valeur du rang moyen est très corrélée avec son écart type. (Plus le rang moyen est faible, plus son écart type est faible).

En ce qui concerne les « Term Frequency », la « Term Frequency » « Binary » fonctionne très mal (c'est la **pire**). Pour rappel, « Binary » signifie que si le terme est présent dans le corpus d'articles d'une maladie, alors TF = 1.0. Ensuite la « Term Frequency » « MinMaxNorm » fonctionne très bien (c'est la **meilleure**). Enfin la « Term Frequency » « RawCount » fonctionne très bien aussi (Juste derrière « MinMaxNorm »). Pour rappel, « RawCount » est le nombre d'occurrences d'une anomalie phénotypique dans le corpus d'une maladie et « MinMaxNorm » est une normalisation de « RawCount ».

En ce qui concerne les IDF, les IDF « IDF_Smooth_NbDisease_i », « IDF_Classic_NbDisease_i » et « Prob_IDF_SumOfMinMaxNorm_i » fonctionnent très bien (« Prob_IDF_SumOfMinMaxNorm_i » est la **meilleure**). L'IDF « Unary » fonctionne très bien curieusement (pour rappel, Unary veut dire qu'on remplace le terme IDF par 1.0). Enfin, on remarquera que les IDF « Inverse_SumOfMinMaxNorm_i » et « Prob_IDF_NbDisease_i » fonctionnent mal (« Prob_IDF_NbDisease_i » est la **pire**).

Pour connaître les formules des différentes versions dont on vient de parler, il suffit de se référer aux Tableaux 8 et 9.

Cependant, si au lieu de regarder ces deux paramètres individuellement, on regarde la combinaison des deux, on peut alors faire les deux remarques suivantes :

- Le point aux coordonnées (Binary, IDF_Classic_SumOfMinMaxNorm_i) représente **le pire** rang moyen ($RMVP = 453.57$)
- Le point aux coordonnées (MinMaxNorm, IDF_Smooth_NbDisease_i) représente **le meilleur** rang moyen ($RMVP = 151.72$)

Si l'on se réfère à la seconde remarque et que l'on considère que le point ayant le rang moyen le plus faible représente la meilleure combinaison (TF, IDF), alors, en utilisant les Tableaux 8 et 9, il semble que le meilleur choix possible pour la formule de TF-IDF à utiliser est la suivante :

$$\boldsymbol{TF - IDF(t, m, M) = \frac{f_{t,m}}{max_{\{t' \in M\}} f_{t',m}} * log\left(1 + \frac{|M|}{f_{inter1,t}}\right)} \qquad (16)$$

où $\boldsymbol{M}$ représente l'ensemble des maladies rares, $\boldsymbol{m}$ une maladie rare et $\boldsymbol{t}$ une anomalie phénotypique.

Supposons que nous utilisions la formule 16 pour calculer la valeur d'importance de nos anomalies extraites. Si nos véritables anomalies se trouvent en moyenne dans les 151.72 premières anomalies classées par ordre décroissant de valeur d'importance (avec un écart type de 157.43, $EcartTypeRangMoyen = 157.43$), quelle proportion cela représente-t-il parmi le nombre d'anomalies extraites en moyenne par l'approche par dictionnaire ? Notre fichier de résultat nous révèle que ce nombre moyen d'anomalies extraites par maladie est de 694.77 anomalies ($\mathrm{NbMoyenAnomalies} = 694.77$). On fait alors le constat suivant : en utilisant la version de TF-IDF ci-dessus, les véritables anomalies se trouvent, en moyenne, aux 21.83% ($RMVP \ / \ NbMoyenAnomalies = 151.72 \ / \ 694.77 = 21.83\%$) de l'ensemble des anomalies extraites avec un écart type de 22.66% ($EcartTypeRangMoyen/NbMoyenAnomalies = 157.43/694.77 = 22.66\%$). Il semble donc que grâce à l'utilisation de cette version, les véritables anomalies se retrouvent souvent dans la première moitié de ce qui est extrait ; cela montre donc que notre manière de calculer la valeur d'importance d'une

anomalie d'une maladie a tendance à favoriser les véritables anomalies plutôt que les anomalies en surplus.

4.6.4.3 CHOIX DU SEUIL

Grâce à l'analyse faite du graphe précédent, nous avons pu faire notre choix concernant la formule à utiliser pour calculer notre valeur d'importance pour nos anomalies phénotypiques. Cet élément étant décidé, il reste le seuil à trouver comme expliqué en 4.5.4. La visualisation qui nous intéresse est alors celle de la Figure 14 qui est paramétrée pour visualiser toutes les valeurs de seuil entre 0 et 0.17 avec un pas de 0.0005.

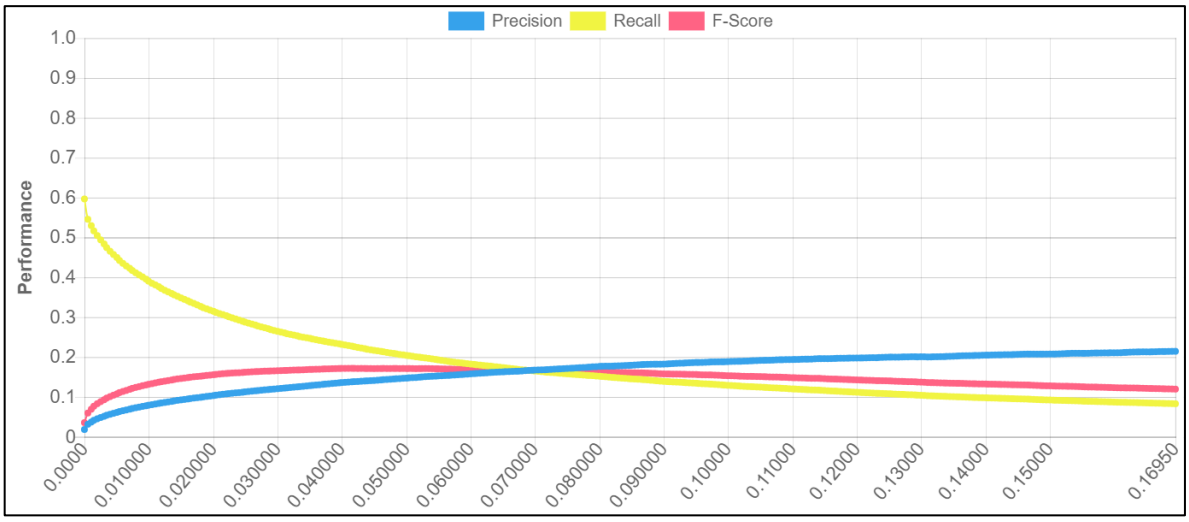

FIGURE 14: VISUALISATION DES FICHIERS "METARESULTSWEIGHT"

L'important à noter est que 3 mesures sont représentées en fonction du seuil choisi :

- En bleu, la précision
- En jaune, le rappel
- En rose, le F-Score

On peut voir que pendant que l'on constate une diminution du rappel, on remarque une augmentation de la précision. D'abord, le fait que la précision augmente signifie que le nombre de faux positifs baisse plus rapidement que le nombre de vrais positifs. Cela signifie que lorsqu'on augmente le seuil, parmi les anomalies éliminées, on élimine plus d'anomalies en surplus que de véritables anomalies et c'est une bonne chose, le contraire aurait posé un problème.

Quant au rappel, sa chute est également tout à fait attendue. En effet, le nombre de vrais positifs va baisser avec l'augmentation du seuil tandis que le nombre de faux négatifs va rester constant, ce qui mathématiquement fait baisser le rappel. La Figure 15 illustre le phénomène, avec, en vert, la courbe du rappel, en rouge, la courbe d'évolution du nombre de vrais positifs et un coefficient a représentant le nombre de faux négatifs. Cette figure est également disponible sur https://www.desmos.com/calculator/ebuzrpdurh si l'on souhaite faire évoluer les paramètres.

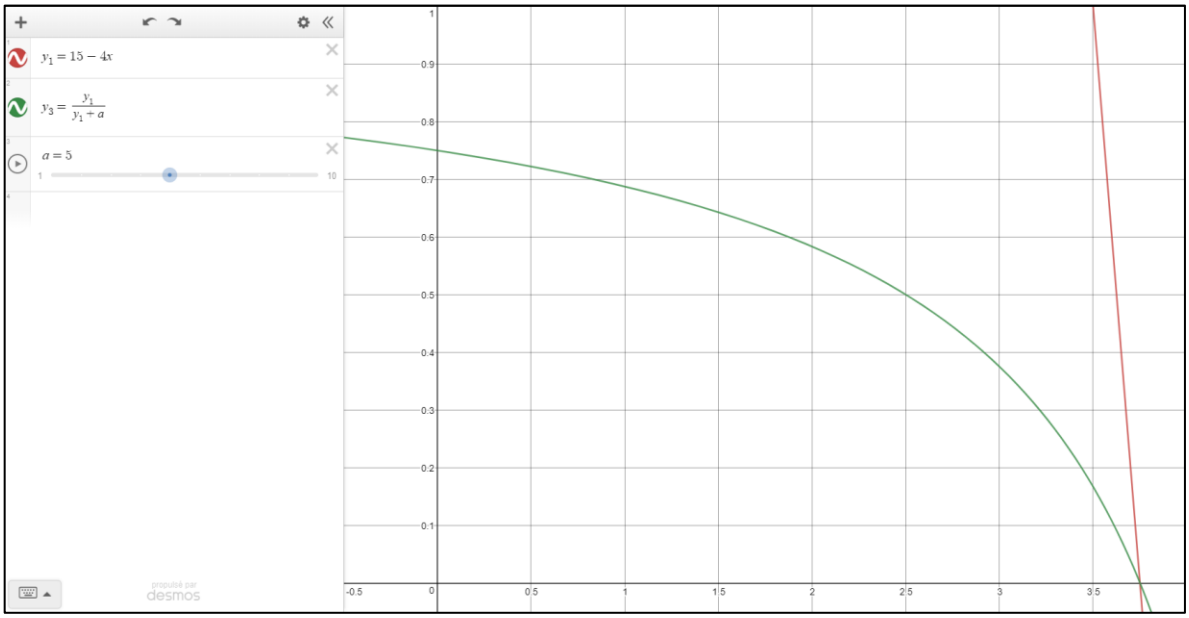

FIGURE 15: ÉVOLUTION DU RAPPEL EN FONCTION DE LA DIMINUTION DU NOMBRE DE VRAIS POSITIFS (NOMBRE DE FAUX NEGATIF=5)

Ensuite, on peut également observer que le F-Score monte puis baisse et que son maximum est proche du point où les courbes de précision et de rappel se croisent. Encore une fois, c'est attendu, le F-score étant une moyenne harmonique de la précision et du rappel.

Mais ce qui nous intéresse est surtout de savoir pour quelle valeur de seuil, le F-score est maximal. Notre outil de visualisation nous permet alors de dire la chose suivante : le meilleur F-score maximal atteint est de **17,17 %** pour une valeur de seuil de **0.0415**. Nous considérerons donc cette valeur de seuil comme celle qu'il faut utiliser pour obtenir la meilleure extraction possible avec notre approche.

4.7 CONCLUSION

Dans ce chapitre, nous avons expliqué les différentes contributions du présent travail de recherche. Tout d'abord, nous avons expliqué nos choix par rapport aux éléments extraits des publications scientifiques, et avons choisi d'extraire les anomalies phénotypiques des maladies rares.

Ensuite, nous avons justifié nos choix concernant nos sources de données. Nous avons choisi de baser notre approche sur les données d'Orphanet et d'OrphaData pour avoir de l'information sur les maladies rares. En ce qui concerne les publications, notre choix s'est orienté vers l'utilisation de la base de données PMC, dont nous avons expliqué le paramétrage. Nous avons également discuté de manière dont le stockage des données serait réalisé.

Troisièmement, nous avons introduit les différents algorithmes qui sont utilisés dans notre approche. Nous avons notamment introduit notre version modifiée de TF-IDF comme mesure d'importance de nos anomalies phénotypiques pour les maladies rares.

Quatrièmement, nous avons décrit le fonctionnement de l'outil développé concernant la NER par dictionnaire, la recherche de la meilleure combinaison de TF-IDF et du meilleur seuil. Nous donnons également des détails sur l'implémentation de celui-ci.

Enfin, nous avons expliqué notre phase d'expérimentation en commençant par notre choix sur les mesures de performances, le format des résultats et de l'outil de visualisation. Finalement, nous exposons les différents résultats obtenus par notre approche. Le choix final pour l'extraction des anomalies phénotypiques des maladies rares est alors de la NER par dictionnaire, dont les anomalies extraites sont filtrées de la manière suivante : après avoir calculé la valeur d'importance de chaque anomalie selon l'Équation 16, les anomalies ayant une valeur d'importance supérieure au seuil de **0.0415** sont conservées. Les anomalies ne respectant pas cette condition sont considérées comme étant non significatives.

CHAPITRE 5

CONCLUSIONS

Cette section présente les conclusions du mémoire. Nous commençons par faire la revue de l'ensemble des contributions, nous parlons ensuite des limites rencontrées ainsi que des travaux futurs envisagés. Nous finissons sur une conclusion personnelle de ce travail de recherche.

5.1 REVUE DES CONTRIBUTIONS

Les contributions apportées par ce travail de recherche sont multiples, elles sont divisées en deux grandes parties : contribution théorique et contribution pratique.

Au niveau théorique, un nouvel algorithme pour la NER pour les anomalies phénotypiques a été créé. Cet algorithme reprend les bases de TF-IDF qui est déjà très utilisé dans le « text mining » en remplaçant le concept de document par celui des maladies rares. À cet algorithme s'ajoute l'utilisation d'un seuil d'élimination des anomalies en surplus. Les performances de l'approche envisagée sont clairement insuffisantes, nous en parlerons plus tard dans les sections concernant les limites et les travaux futurs, mais elle n'en reste pas moins une tentative d'innovation dans le domaine de la NER.

Au niveau pratique, un système développé en quatre parties distinctes. Premièrement, le système permet de récolter les informations sur les maladies rares ainsi que les publications qui leur sont liées puis de stocker l'ensemble dans une base de données MongoDB. La deuxième partie du système s'occupe, quant à elle, de réaliser l'extraction des anomalies phénotypiques des publications ainsi que de l'attribution de la valeur d'importance à chacune d'entre elles et du stockage dans la base de données. Ensuite, le système est capable de s'occuper de l'évaluation et plus précisément de la production de fichiers de performances. Ces fichiers seront ensuite visualisables sous forme de graphiques et de statistiques grâce à l'outil de visualisation qui est la quatrième partie du système développé.

5.2 LIMITES

Lors de la création de notre système d'extraction, nous avons constaté que nous étions limités par certains éléments. La première chose est que ce travail de recherche repose sur un certain nombre d'hypothèses et parmi ces hypothèses, nous pensons que l'une d'entre elles est invalide ou du moins pas totalement vraie.

Si on prend l'hypothèse H1 qui suppose la suffisance de la quantité d'informations sur les maladies rares dans les publications scientifiques, nous nous sommes rendu compte qu'elle n'est pas respectée pour différentes raisons. Premièrement, il existe des maladies qui se trouvent être si rares que la littérature scientifique à leurs propos est quasi inexistante. Et ce ne serait pas lié à notre source pour récupérer des publications, mais lié simplement à la rareté de la maladie elle-même. Deuxièmement, notre source de publication, à savoir PMC (que nous avions choisi principalement parce qu'elle propose des publications avec leur texte complet et parce qu'elle propose le plus de publications parmi les différentes sources envisagées) ne semble pas contenir suffisamment de publications en ce qui concerne les maladies rares. Pour appuyer ces deux points, on peut se référer au graphique de la Figure 12. Si l'on regarde l'évolution du nombre de maladies en fonction du nombre de publications, on remarque alors que sur les 2572 maladies qui ont pu être évaluées, 1073 possèdent entre 0 et 100 publications, soit 41.72 % des maladies. Nous ne pouvons pas dire que c'est une preuve que le nombre de publications est insuffisant pour un grand nombre de maladies rares, mais nous pensons que cela témoigne de la difficulté que l'on peut avoir à extraire automatiquement des informations sur les maladies rares à partir de publications.

Ensuite, au vu des résultats obtenus pour extraire les anomalies phénotypiques avec notre approche, à savoir de la NER par dictionnaire et un filtrage par seuil sur la valeur d'importance calculée à partir de notre TF-IDF modifié (17.17 %) et des performances que l'on peut obtenir avec de la NER par dictionnaire simple (3.55 %), on comprend facilement que ce filtrage permet d'améliorer les performances de la NER par dictionnaire, mais que c'est encore bien insuffisant pour espérer faire une extraction de qualité.

5.3 TRAVAUX FUTURS

Au vu des difficultés et des limites émises dans la section précédente, il est assez facile d'envisager différentes pistes pour les travaux futurs. La première chose à considérer serait la possibilité d'améliorer les performances par la considération d'autres critères statistiques. Par exemple, l'emploi d'analyse syntaxique des phrases contenant des termes faisant référence à des anomalies phénotypiques et l'utilisation des résultats de cette analyse pourrait avoir un effet significatif sur les performances.

Un deuxième élément à considérer est l'extension du dictionnaire d'anomalies phénotypiques pour augmenter la quantité de symptômes reconnus. En effet, la liste de symptômes fournie par HPO n'est pas exhaustive et pourrait être étendue par des méthodes d'apprentissage automatique ou semi-automatique.

Enfin, on peut également imaginer d'extraire d'autres types d'éléments que les symptômes dans les publications scientifiques comme des médicaments, des traitements, les causes potentielles, les maladies liées ; ces éléments, étant des informations extrêmement intéressantes pour le grand public, pourraient ainsi être mis à disposition de manière complètement automatique.

5.4 CONCLUSION PERSONNELLE

J'aimerais terminer la rédaction de ce mémoire avec une conclusion plus personnelle sur ce travail de recherche. Ce travail aura été ma première véritable expérience dans le domaine de la recherche scientifique, il aura nécessité beaucoup d'heures de travail et une constante autoformation dans le domaine du « text mining ». Cependant, je ne regrette rien de tout cela, je me sens grandi, plus compétent, plus sûr de moi et les connaissances que j'ai pu acquérir pendant près d'un an sont inestimables pour moi.

Après une première expérience aussi positive que celle-ci, continuer à travailler dans le domaine de la recherche en informatique semble me correspondre au mieux et j'ai hâte de pouvoir mettre à profit mes efforts pour aider à pousser les avancées scientifiques le plus loin possible.

Pour finir, je voudrais également remercier ma famille et mes amis pour leur soutien ainsi que l'ensemble de mes collègues de travail au laboratoire qui m'ont guidé d'une manière ou d'une autre pour mener ce projet de recherche au mieux.

RÉFÉRENCES

Aggarwal, C. C., & Han, J. (2014). *Frequent Pattern Mining*. Springer Publishing Company, Incorporated.

Alias-i. (2008). LingPipe. Repéré le 03/03/2018, à http://alias-i.com/lingpipe/

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

Bellman, R. (2013). *Dynamic programming*. Courier Corporation.

Breckbaldwin. (2009, 2009/10/14/). Coding Chunkers as Taggers: IO, BIO, BMEWO, and BMEWO+. Repéré à https://lingpipe-blog.com/2009/10/14/coding-chunkers-as-taggers-io-bio-bmewo-and-bmewo

Bui, Q.-C., & Sloot, P. M. A. (2012). A robust approach to extract biomedical events from literature. *Bioinformatics*, *28*(20), 2654-2661. doi: 10.1093/bioinformatics/bts487

Burges, C. J. (2010). From ranknet to lambdarank to lambdamart: An overview. *Learning*, *11*(23-581), 81.

Cadiet. (2017). *L'open data des décisions de justice*. Repéré à http://www.justice.gouv.fr/publication/open_data_rapport.pdf

Cha, J., Kim, J., Yeu, Y., & Park, S. (2016). *A method for obtaining rich data from PubMed using SVM*. Communication présentée au Proceedings of the 31st Annual ACM Symposium on Applied Computing, Pisa, Italy. Repéré à http://dl.acm.org/citation.cfm?doid=2851613.2851866

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51-89. doi: doi:10.1002/aris.1440370103

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in bioinformatics*, *6*(1), 57-71.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273-297. doi: 10.1023/a:1022627411411

Deng, J., Dong, W., Socher, R., Li, L., Kai, L., & Li, F.-F. (2009, 20-25 June 2009). *ImageNet: A large-scale hierarchical image database*. Communication présentée au 2009 IEEE Conference on Computer Vision and Pattern Recognition,

Deshpande, B. (2012, 2017/10/31/). 3 ways to use text mining with RapidMiner to juice up your job search. Repéré le 31/10/2017, à http://www.simafore.com/blog/bid/111839/3-ways-to-use-text-mining-with-RapidMiner-to-juice-up-your-job-search

Doughty, E., Kertesz-Farkas, A., Bodenreider, O., Thompson, G., Adadey, A., Peterson, T., & Kann, M. G. (2011). Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. *Bioinformatics*, *27*(3), 408-415. doi: 10.1093/bioinformatics/btq667

Erdogan, H. (2010). *Sequence labeling: Generative and discriminative approaches*.

Europe P.M.C. Consortium. (2015). Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic acids research*, *43*(Database issue), D1042-1048. doi: 10.1093/nar/gku1061

Fox, C. (1989). A stop list for general text. *SIGIR Forum*, *24*(1-2), 19-21. doi: 10.1145/378881.378888

Goldberg, Y., & Levy, O. (2014). word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Google. (2004). Google Scholar. Repéré le 05/05/2018, à https://scholar.google.ca/

Heeringa, W. J. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. Citeseer.

Hirschman, L., Yeh, A., Blaschke, C., & Valencia, A. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC bioinformatics*, *6 Suppl 1*(Suppl 1), S1-S1. doi: 10.1186/1471-2105-6-S1-S1

Holat, P., Tomeh, N., Charnois, T., Battistelli, D., Jaulent, M.-C., & Métivier, J.-P. (2016). *Weakly-Supervised Symptom Recognition for Rare Diseases in Biomedical Text*. Communication présentée au International Symposium on Intelligent Data Analysis, Stockholm.

Inserm. (1999). Orphanet: À propos des maladies rares. Repéré le 03/03/2018, à http://www.orpha.net/consor/cgi-bin/Education_AboutRareDiseases.php?lng=FR

Kao, A., & Poteet, S. R. (2007). *Natural language processing and text mining*. Springer Science & Business Media.

Köhler, S., Vasilevsky, N. A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., ... Robinson, P. N. (2017). The Human Phenotype Ontology in 2017. *Nucleic Acids Research*, *45*(D1), D865-D876. doi: 10.1093/nar/gkw1039

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Li, A., Zang, Q., Sun, D., & Wang, M. (2016). A text feature-based approach for literature mining of lncRNA–protein interactions. *Neurocomputing*, *206*, 73-80. doi: 10.1016/j.neucom.2015.11.110

Lovins, J. B. (1968). Development of a stemming algorithm. *Mech. Translat. & Comp. Linguistics*, *11*(1-2), 22-31.

Mahmood, A. S. M. A., Wu, T.-J., Mazumder, R., & Vijay-Shanker, K. (2016). DiMeX: A Text Mining System for Mutation-Disease Association Extraction. *PLoS ONE*, *11*(4), 1-26. doi: 10.1371/journal.pone.0152725

Martin, L., Battistelli, D., & Charnois, T. (2014, 2014-06-22). *Symptom recognition issue*. Communication présentée au 13th workshop on Biomedical Natural Language Processing

(BioNLP 2014), Baltimore, United States. Repéré à https://halshs.archives-ouvertes.fr/halshs-01727094

Merz, T. (2018, 2018/07/11/). TET. Repéré à https://www.pdflib.com/download/tet

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, *30*(1), 3-26. doi: https://doi.org/10.1075/li.30.1.03nad

Neimke. (2003). Regular Expression Library. Repéré le 21/12/2017, à http://regexlib.com/REDetails.aspx?regexp_id=356

Orphanet. (2015, 2015/01/29/). La bêta-thalassémie. Repéré à https://www.orpha.net/data/patho/Pub/fr/BetaThalassemie-FRfrPub51.pdf

Orphanet. (2016, 2016/03/30/). Rapport d'Activité 2016 d'Orphanet. Repéré à https://www.orpha.net/orphacom/cahiers/docs/FR/Rapport_activite_2016.pdf

Porter, M. F. (2001). Snowball: A language for stemming algorithms. Repéré à http://snowball.tartarus.org/

Rabiner, L. R., & Juang, B.-H. (1986). An introduction to hidden Markov models. *IEEE assp magazine*, *3*(1), 4-16.

Sailaja, V. N., Padmasree, L., & Mangathayaru, N. (2016). Survey of Text Mining Techniques Challenges and their Applications. *International Journal of Computer Applications*, *146*(11), 30-35.

Sayers, E. (2010). A General Introduction to the E-utilities. *Entrez Programming Utilities Help [Internet]. Bethesda: National Center for Biotechnology Information*.

Schulman, P., Castellon, C., & Seligman, M. E. (1989). Assessing explanatory style: the content analysis of verbatim explanations and the Attributional Style Questionnaire. *Behav Res Ther*, *27*(5), 505-512.

Singhal, A., Simmons, M., & Lu, Z. (2016). Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine. *PLoS Computational Biology*, *12*(11), 1-19. doi: 10.1371/journal.pcbi.1005017

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *28*(1), 11-21.

Stephen, R. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, *60*(5), 503-520. doi: doi:10.1108/00220410410560582

Sutton, C., & McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, *4*(4), 267-373.

Tanabe, L., & Wilbur, W. J. (2002). Tagging gene and protein names in biomedical text. *Bioinformatics*, *18*(8), 1124-1132. doi: 10.1093/bioinformatics/18.8.1124

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company. Repéré à https://books.google.co.uk/books?id=UT9dAAAAIAAJ

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*(2), 260-269. doi: 10.1109/TIT.1967.1054010

Voutilainen, A. (2003). Part-of-speech tagging. *The Oxford handbook of computational linguistics*, 219-232.

Vukotic, V., Claveau, V., & Raymond, C. (2015, 2015-06-22). *IRISA at DeFT 2015: Supervised and Unsupervised Methods in Sentiment Analysis*. Communication présentée au DeFT, Défi Fouille de Texte, joint à la conférence TALN 2015, Caen, France. Repéré à https://hal.archives-ouvertes.fr/hal-01226528

Wikipedia. (2017). Publication Scientifique. Repéré le 27/03/2017, à https://fr.wikipedia.org/wiki/Publication_scientifique

Yeh, A. S., Hirschman, L., & Morgan, A. A. (2003). Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics*, *19*(suppl_1), i331-i339.

Yifan, P., Chih-Hsuan, W., & Zhiyong, L. (2016). Improving chemical disease relation extraction with rich features and weakly labeled data. *Journal of Cheminformatics*, *8*, 1-12. doi: 10.1186/s13321-016-0165-z

Yvon, F. (2016, 18/04/2017). Une petite introduction au Traitement Automatique des Langues Naturelles. Repéré le 03/03/2018, à https://perso.limsi.fr/anne/coursM2R/intro.pdf